Lecture Notes in Mathematics Number 58

Infinitely Large Neural Networks

Ernest K. Ryu

Research Institute of Mathematics
Seoul National University, Seoul 08826, Korea

Lecture Notes Series in Mathematics

58

Infinitely Large Neural Networks

Ernest K. Ryu

Published by the
Research Institute of Mathematics
Seoul National University, Seoul 08826, Korea
Lecture Notes in Mathematics

Research Institute of Mathematics
Seoul National University

Ernest K. Ryu
Department of Mathematical Sciences and Resesarch Institute of Mathematics
Seoul National University
Seoul 08826
Korea

2020 Mathematics Subject Classification. 46S99, 68T99, 90C25.

Printed in Korea
2023. 1. 10.

Infinitely Large Neural Networks

Ernest K. Ryu

Department of Mathematics, Seoul National University

January, 2023

ABSTRACT

During the week of December 20-24, 2004, the author is one of two principal lecturers at the Winter School 2018 of Gauss-Hilbert Theory. In this lecture I attempt to set forth some of the recent developments that had taken place in Gauss-Hilbert Theory.

Contents

Chapter 1

Universal approximation theory for wide neural networks

1.1 Cybenko's proof

Let $\Omega \subset \mathbb{R}^d$ be compact. Let $\sigma\colon \mathbb{R} \to \mathbb{R}$. Let f_θ represents width-N 2-layer neural networks:

$$f_\theta(x) = \sum_{i=1}^{N} u_i \sigma(a_i^\intercal x + b_i), \tag{1.1}$$

where $\theta \in \Theta^{(N)}$ and

$$\Theta^{(N)} = \{(a_1, \dots, a_N, b_1, \dots, b_N, u_1, \dots, u_N) \,|\, a_1, \dots, a_N \in \mathbb{R}^d,\, b_1, \dots, b_N, u_1, \dots, u_N \in \mathbb{R}\}.$$

To clarify, $\Theta^{(N)} \cong \mathbb{R}^{dN+2N}$.

Theorem 1. *Let $\Omega \subset \mathbb{R}^d$ be compact. Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying*

$$\lim_{r\to-\infty} \sigma(r) = 0, \qquad \lim_{r\to\infty} \sigma(r) = 1.$$

The class of functions

$$\bigcup_{N\in\mathbb{N}} \{f_\theta\}_{\theta\in\Theta^{(N)}}$$

is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$, i.e.,

$$\text{closure}(\text{span}(\{\sigma(a^\intercal x + b)\}_{a\in\mathbb{R}^d, b\in\mathbb{R}})) = (\mathcal{C}(\Omega), \|\cdot\|_\infty).$$

The consequence of this theorem is that for any $f_\star \in \mathcal{C}(\Omega)$ and $\varepsilon > 0$, there exists a large enough N and network parameter $\theta \in \Theta^{(N)}$ such that

$$\sup_{x\in\Omega} |f_\theta(x) - f_\star| < \varepsilon.$$

Smaller ε will likely require larger N, but this theorem or its proof will not allow us to make any quantitative claims. Moreover, this is an existence result; it does not tell us how to find N and $\theta \in \Theta^{(N)}$.

Pseudo-proof in dimension 1. Let $d = 1$. Note that

$$\sigma\left(\frac{a}{\delta}x - \frac{b}{\delta}\right) \to \mathbf{1}_{[b/a,\infty)}$$

as $\delta \to 0$, i.e., for small δ, we have a smooth approximation of the step function. (Note for $x = \frac{b}{a}$, we do not have convergence. Remember that this is not a real proof.) Then

$$\mathbf{1}_{[t_0,t_1]} \approx \sigma\left(\frac{1}{\delta}x - \frac{t_0}{\delta}\right) - \sigma\left(\frac{1}{\delta}x - \frac{t_1}{\delta}\right)$$

for small $\delta > 0$.

Given a smooth function $f_\star$, find a piecewise constant approximation to it. Then form a smooth approximation of the piecewise constant approximation. □

The actual proof of Theorem 1 will be done in two steps, with the following Lemmas 1 and 2. We say σ is *discriminatory* if

$$\left[\mu \in \mathcal{M}(\Omega) \text{ such that } \int_\Omega \sigma(a^\intercal x + b)\, d\mu(x) = 0,\ \forall\, a \in \mathbb{R}^d,\, b \in \mathbb{R}\right] \quad \Rightarrow \quad \mu = 0.$$

In functional analysis, one often views an object in a primal and a dual way. In the primal view, we view $\mu \in \mathcal{M}(\Omega)$ as mapping that assigns a "volume" to any measurable set. In the dual view, we instead view the action of $L_\mu \colon C(\Omega) \to \mathbb{R}$ defined by

$$L_\mu[f] = \int_\Omega f(x)\, d\mu(x).$$

Under this view, σ is discrminatory if the fact that $L_\mu[\sigma(a^\intercal \cdot + b)] = 0$ for all a and b implies that $L_\mu[f] = 0$ for all f, i.e., to determine whether $L_\mu = 0$, it is sufficient to check all inputs of the form $\sigma(a^\intercal \cdot + b)$.

We quickly provide some non-examples. Let $d = 1$ and $\Omega = [-10, 10]$. Then $\sigma(x) = 1$ is not discriminatory since

$$\mu = -\delta_{-1} + \delta_1 \neq 0,$$

where δ_r is the Dirac delta measure centered at $r \in \mathbb{R}$, satisfies

$$\int_\Omega \sigma(ax + b) = 0, \qquad \forall a \in \mathbb{R}^d,\, b \in \mathbb{R}.$$

Likewise, $\sigma(x) = x$ is not discriminatory since

$$\mu = \frac{1}{2}\delta_{-1} - \delta_0 + \frac{1}{2}\delta_1 \neq 0$$

satisfies

$$\int_\Omega \sigma(ax + b) = 0, \qquad \forall a \in \mathbb{R}^d,\, b \in \mathbb{R}.$$

One can show that all polynomials are not discriminatory.

Lemma 1. *Let $\Omega \subseteq \mathbb{R}^d$. If σ is discriminatory, then $\{f_\theta\}$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.*

Proof. Let $\mathcal{S} = \operatorname{span}(\{\sigma(a^\intercal x + b)\}_{a\in\mathbb{R}^d,\, b\in\mathbb{R}}) \subseteq \mathcal{C}(\Omega)$, and let $\overline{\mathcal{S}}$ be its closure in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$. Assume for contradiction that $\overline{\mathcal{S}} \neq \mathcal{C}(\Omega)$. Then pick $g \neq 0$, $g \in \mathcal{C}(\Omega)\backslash\overline{\mathcal{S}}$ and define the linear form L on $\overline{\mathcal{S}} \oplus \operatorname{span}(g)$ as

$$L[s + \lambda g] = \lambda, \qquad \forall\, s \in \overline{\mathcal{S}},\, \lambda \in \mathbb{R}.$$

This makes L a bounded linear operator[1] such that $L = 0$ on $\overline{\mathcal{S}}$ but $L \neq 0$. Using the Hahn–Banach theorem, we can extend L to $\overline{L}\colon \mathcal{C}(\Omega) \to \mathbb{R}$ such that $\overline{L}$ is bounded and linear. Since $\overline{L} \in \mathcal{C}(\Omega)^* \cong \mathcal{M}(\Omega)$, there exists $\mu_{\overline{L}} \in \mathcal{M}(\Omega)$ such that $\overline{L}(h) = \int_\Omega h\, d\mu_{\overline{L}}$. However, $\overline{L} = 0$ on $\overline{\mathcal{S}}$, so

$$\int \sigma(a^\intercal x + b)\, d\mu_{\overline{L}}(x) = 0$$

for all a, b and $\mu_{\overline{L}} = 0$ and $\overline{L} = 0$ by the hypothesis. $\overline{L} = 0$ contradicts the construction $L[s + \lambda g] = \lambda$, so we conclude $\overline{\mathcal{S}} = \mathcal{C}(\Omega)$. □

Lemma 2. *A σ satisfying the condition of Theorem 1 is discriminatory.*

Proof. Define

$$\varphi_{a,b} = \sigma(a^\intercal x + b)$$

and

$$\begin{aligned} H_{a,b} &= \{x \mid a^\intercal x + b > 0\} && \text{(open half-space defined by } (a, b)) \\ \partial H_{a,b} &= \{x \mid a^\intercal x + b = 0\} && \text{(boundary of the half space)} \end{aligned}$$

[1] In the current setup, $\operatorname{dist}(g, \overline{\mathcal{S}}) > 0$, since $\overline{\mathcal{S}}$ is closed, and $\|s + \lambda g\| = |\lambda|\|(1/\lambda)s + g\| \geq |\lambda| \operatorname{dist}(g, \overline{\mathcal{S}})$. So $|L[s + \lambda g]| = |\lambda| \leq \frac{1}{\operatorname{dist}(g,\overline{\mathcal{S}})}\|s + \lambda g\|$, and L is bounded.

for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Then

$$\varphi_{\frac{a}{\delta},\frac{b}{\delta}+t}(x) = \sigma\left(\frac{a^\intercal x + b}{\delta} + t\right) \overset{\delta\to 0}{\to} \gamma_t = \begin{cases} 1 & \text{if } x \in H_{a,b} \\ \sigma(t) & \text{if } x \in \partial H_{a,b} \\ 0 & \text{otherwise} \end{cases}$$

pointwise. By the Lebesgue dominated convergence theorem (since σ is bounded)

$$\int_\Omega \varphi_{a/\delta,b/\delta+t}(x)\, d\mu(x) \to \int_\Omega \gamma_t(x)\, d\mu(x) = \sigma(t)\mu(\partial H_{a,b}) + \mu(H_{a,b}).$$

Since the question is whether σ is discriminatory, consider the scenario where all of these integrals vanish. Then $\sigma(t)\mu(\partial H_{a,b}) + \mu(H_{a,b}) = 0$ for all $t \in \mathbb{R}$, $a \in \mathbb{R}^d$, and $b \in \mathbb{R}$. Since σ is not a constant function, this implies $\mu(\partial H_{a,b}) = \mu(H_{a,b}) = 0$. If this implies that $\mu = 0$, then σ is discriminatory.

We now show $\mu = 0$. For $a \in \mathbb{R}^d$, consider

$$F_{a,\mu}[h] = \int_\Omega h(a^\intercal x)\, d\mu(x),$$

which is linear. We have

$$F_{a,\mu}[\mathbf{1}_{[-b,\infty)}] = \int_\Omega \mathbf{1}_{[-b,\infty)}(a^\intercal x)\, d\mu(x) = \mu(\partial H_{a,b}) + \mu(H_{a,b}) = 0.$$

We define *step functions* to be functions of the form

$$\sum_{i=1}^{N} c_i \mathbf{1}_{[t_i,t_{i+1})}$$

with $N \in \mathbb{N}$, $t_1 < t_2 < \cdots < t_N$, and $c_1, \ldots, c_N \in \mathbb{R}$. By linearity,

$$F_{a,\mu}[h] = 0$$

for all step functions h. There exists a sequence of step functions $h_1, h_2, \ldots$ such that

$$|h_i(x)| \le |\sin(x)|, \qquad h_i(x) \to \sin(x), \qquad \forall\, x \in \mathbb{R}.$$

By the Lebesgue dominated convergence theorem,

$$\int_\Omega \sin(a^\intercal x)\, d\mu(x) = 0.$$

We can make the same argument with $\cos(x)$. Combining the two cases, we get

$$\hat{\mu}(a) = \int_\Omega e^{ia^\intercal x}\, d\mu(x) = 0.$$

Since the Fourier transform of μ is zero, we conclude $\mu = 0$. □

But what if the function we wish to approximate is discontinuous? For example, what if we wish to approximate a function $f_\star \colon \Omega \to \{1, \dots, k\}$. While one can approximate continuous functions with discontinuous functions, one cannot approximate continuous functions with discontinuous ones in the $\|\cdot\|_\infty$-norm.

However, we can have the continuous function approximate the discontinuous function is most of the domain.

Theorem 2 (Lusin's theorem)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be compact. Let $f \colon \Omega \to \mathbb{R}$ be a measurable function. For any $\varepsilon > 0$, there exists a continuous function $f_\varepsilon \colon \Omega \to \mathbb{R}$ and $\Omega' \subseteq \Omega$ such that $\mathrm{Vol}(\Omega \backslash \Omega') < \varepsilon$ and such that*

$$f(x) = f_\varepsilon(x), \qquad \forall\, x \in \Omega'.$$

(Here, Vol denotes the "volume" defined by the Lebesgue measure.)

Theorem 3. *Consider the setup of Theorem 1. Let $f_\star \colon \Omega \to \{1, \dots, k\}$ be a (measurable) decision function. For any $\varepsilon > 0$, there exists an f_θ and $\Omega' \subseteq \Omega$ such that $\mathrm{Vol}(\Omega \backslash \Omega') < \varepsilon$ and*

$$|f_\star(x) - f_\theta(x)| < \varepsilon, \qquad \forall\, x \in \Omega'.$$

Proof. By Lusin's theorem, we find f_ε and Ω' such that $f_\varepsilon = f_\star$ on Ω'. Then we appeal to Theorem 1 to find an f_θ such that $f_\theta \approx f_\varepsilon$ on Ω. Then $f_\theta \approx f_\star$ on Ω'. □

1.2 Applications of Stone–Weierstrass

Theorem 4 (Stone–Weierstrass)**.** *Let $\Omega \subset \mathbb{R}^d$ be compact. Let $\mathcal{F} \subseteq (\mathcal{C}(\Omega), \|\cdot\|_\infty)$ be a subalgebra that contains the non-zero constant function. Then $\mathcal{F}$ is dense if and only if for any distinct $x, y \in \Omega$ there exists an $f \in \mathcal{F}$ such that*

$$f(x) \neq f(y).$$

For the following class of functions (not 2-layer neural networks, but implementable) the Stone–Weierstrass theorem immediately applies.

$$f_\theta(x) = \sum_{i=1}^{N} u_i \prod_{j=1}^{M_i} \sigma(a_{ij}^\intercal x + b_{ij}). \tag{1.2}$$

Lemma 3. *The class of functions of the form (1.2) for all $N, M_1, \dots, M_N \in \mathbb{N}$ is an algebra.*

Lemma 4. *Let $\Omega \subseteq \mathbb{R}^d$ be compact. If $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a continuous nonconstant function, then functions of the form* (1.2) *is dense in* $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.

Proof. The statement follows from the Stone–Weierstrass theorem. By the previous lemma, it remains to establish the separation requirement. Since σ is nonconstant, there exists $r_1, r_2 \in \mathbb{R}$ such that $\sigma(r_1) \neq \sigma(r_2)$. Then for any distinct $x, y \in \Omega$, there exists $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$a^\intercal x + b = r_1, \qquad a^\intercal y + b = r_2.$$

Then $\sigma(a^\intercal x + b) \neq \sigma(a^\intercal y + b)$. □

Theorem 5. *Let $\Omega \subset \mathbb{R}^d$ be compact. Let $\sigma = \sin$. The class of 2-layer neural networks* (1.1) *is dense in* $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.

Proof outline. The class of functions of the form (1.2) is dense by Lemma 4. Using the trigonometric identity

$$2\sin(a)\sin(b) = \sin\left(a + b - \frac{\pi}{2}\right) - \sin\left(a - b - \frac{\pi}{2}\right),$$

we can convert functions of the form (1.2) into functions of the form (1.1). Therefore,

$$\{\text{functions of the form (1.2)}\} \subseteq \{\text{functions of the form (1.1)}\},$$

so functions of the form (1.1) is also dense. □

Theorem 6. *Let $\Omega \subseteq \mathbb{R}^d$ be compact. If $\mu \in \mathcal{M}(\Omega)$ such that*

$$\hat{\mu}(a) = \int_\Omega e^{ia^\intercal x}\, d\mu(x) = 0$$

for all $a \in \mathcal{R}^d$. Then $\mu = 0$.

Proof. Homework exercise. □

Next, we will established the following further general universality result.

Theorem 7 (Leshno [2])**.** *Let $\sigma \in \mathcal{C}(\mathbb{R})$ be non-polynomial. Let $\Omega \subset \mathbb{R}^d$ be compact. Then* $\operatorname{span}\{\sigma(a^\intercal \cdot + b) \mid a \in \mathbb{R}^d,\, b \in \mathbb{R}\}$ *is dense in* $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.

The first step of our proof will be to reduce the universality in the d-dimensions to 1-dimension. For any $\sigma\colon \mathbb{R} \to \mathbb{R}$, define

$$\mathcal{S}^1 = \operatorname{span}\{\sigma(s \cdot + t) \mid s \in \mathbb{R},\, t \in \mathbb{R}\}$$

and

$$\mathcal{S}^d = \operatorname{span}\{\sigma(a^\intercal \cdot + b) \mid a \in \mathbb{R}^d,\, b \in \mathbb{R}\}.$$

Lemma 5. *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be continuous[2] such that $\mathcal{S}^1$ is dense in $(\mathcal{C}(K), \|\cdot\|_\infty)$ for any compact $K \subset \mathbb{R}$. Then $\mathcal{S}^d$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$ for any compact $\Omega \subset \mathbb{R}^d$.*

Proof. Since $\operatorname{span}\{\sin(a^\intercal \cdot + b) \mid a \in \mathbb{R}^d,\, b \in \mathbb{R}\}$ is dense in $\mathcal{C}(\Omega)$ by Theorem 5, there exists

$$\left| f_\star(x) - \sum_{i=1}^{N} u_i \sin(a_i^\intercal x + b_i) \right| < \frac{\varepsilon}{2}, \qquad \forall\, x \in \Omega.$$

Let

$$D = \sup_{\substack{x \in \Omega \\ i=1,\dots,N}} |a_i^\intercal x|.$$

Since $\mathcal{S}^1$ is dense in $\mathcal{C}([-D, D])$, there exists

$$\left| u_i \sin(a_i^\intercal x + b_i) - \sum_{j=1}^{M_i} v_{ij} \sigma(s_{ij}(a_i^\intercal x) - t_{ij}) \right| < \frac{\varepsilon}{2N}, \qquad \forall\, i = 1, \dots, N,\, x \in \Omega.$$

By the triangle inequality, we conclude

$$\left| f_\star(x) - \sum_{i=1}^{N} \sum_{j=1}^{M_i} v_{ij} \sigma(s_{ij}(a_i^\intercal x) - t_{ij}) \right| < \varepsilon, \qquad \forall\, x \in \Omega.$$

□

Lemma 6. *$\sigma \in \mathcal{C}^\infty(\mathbb{R})$ is a polynomial of degree at most $k \in \mathbb{N}$ if and only if*

$$\sigma^{(k+1)}(t) = 0, \qquad \forall\, t \in \mathbb{R}.$$

Lemma 7. *Let $\sigma \in \mathcal{C}^\infty(\mathbb{R})$. Let $K \subseteq \mathbb{R}$ be compact. Then*[3]

$$r^k \sigma^{(k)}(t) \in \overline{\mathcal{S}^1}$$

for all $k \in \mathbb{N}$ and $t \in \mathbb{R}$, where the closure is taken in $(\mathcal{C}(K), \|\cdot\|_\infty)$.

[2]The proof really only requires $\sigma\colon \mathbb{R} \to \mathbb{R}$ to be measurable. However, we assume σ is continuous so that $\mathcal{S}^1 \subseteq \mathcal{C}(K)$ and $\mathcal{S}^d \subseteq \mathcal{C}(\Omega)$ for $K \subset \mathbb{R}$ and $\Omega \subset \mathbb{R}^d$.

[3]Since $\sigma\colon \mathbb{R} \to \mathbb{R}$, we allow $t \in \mathbb{R}$, rather than restricting t to the interior of K. The set $K \subseteq \mathbb{R}$ is used for defining the notion of convergence rather than for restricting the input.

Proof. Let $\mathcal{S}^1 = \operatorname{span}\{\sigma(sr+t) \mid s \in \mathbb{R},\, t \in \mathbb{R}\}$. Then

$$r\sigma'(t) = \frac{d}{ds}\sigma(sr+t)\Big|_{s=0} = \lim_{h\to 0} \frac{\sigma(hr+t)-\sigma(t)}{h} \in \overline{\mathcal{S}^1}.$$

We do need to verify that the convergence is uniform on compact K. Since

$$\begin{aligned}\frac{\sigma(hr+t)-\sigma(t)}{h} - r\sigma'(t) &= \frac{r}{h}\int_0^h (\sigma'(\eta r+t)-\sigma'(t))\,d\eta \\ &= \frac{r^2}{h}\int_0^h \int_0^\eta \sigma''(\nu r+t)\,d\nu\,d\eta,\end{aligned}$$

we have

$$\sup_{r\in K}\left|\frac{\sigma(hr+t)-\sigma(t)}{h} - r\sigma'(t)\right| \le h\left(\sup_{r\in K} r^2\right)\left(\sup_{r\in K}|\sigma''(r)|\right) < \infty, \qquad \forall\, h \neq 0.$$

Likewise,

$$r^k\sigma^{(k)}(t) = \frac{d^k}{ds^k}\sigma(rs+t)\Big|_{s=0} \in \overline{\mathcal{S}^1}.$$

□

Corollary 1. *Let $\sigma \in \mathcal{C}^\infty(\mathbb{R})$ and assume σ is not a polynomial. Then $\mathcal{S}^1$ is dense in $(\mathcal{C}(K), \|\cdot\|_\infty)$ for any compact $K \subseteq \mathbb{R}$.*

Proof. For any $k \in \mathbb{N}$, there is a $t \in \mathbb{R}$ such that $\sigma^{(k)}(t) \neq 0$, by Lemma 6, and implies $r^k \in \overline{\mathcal{S}^1}$, by Lemma 7. Since $\overline{\mathcal{S}^1}$ contains all monomials, $\overline{\mathcal{S}^1}$ is dense in $\mathcal{C}(K)$ by Stone–Weierstrass. ($\overline{\mathcal{S}^1}$ is not an algebra, but it contains the subalgebra of polynomials, which is dense in $\mathcal{C}(K)$ by Stone–Weierstrass.) □

Let

$$\Psi(t) = \begin{cases} \exp\left(-\frac{1}{1-t^2}\right) & \text{for } t \in (-1,1) \\ 0 & \text{otherwise} \end{cases}$$

be our smooth "bump" function. Such a function is also referred to as a *mollifier*. Define

$$\phi_\delta(t) = \frac{1}{\delta\int_{\mathbb{R}}\Psi(t)\,dt}\Psi(t/\delta).$$

Then $\phi_\delta \in \mathcal{C}^\infty$ is supported on $[-\delta,\delta]$ and $\int_{\mathbb{R}}\phi_\delta(t)\,dt = 1$. For any $\sigma \in \mathcal{C}(\mathbb{R})$, define

$$\sigma_\delta(r) = \int_{\mathbb{R}} \sigma(r-t)\phi_\delta(t)\,dt.$$

Lemma 8. *For any $\sigma \in \mathcal{C}(\mathbb{R})$ and any compact $K \subseteq \mathbb{R}$,*

$$\sup_{r \in K} |\sigma_\delta(r) - \sigma(r)| \to 0$$

as $\delta \to 0$.

Proof. Define $K' = [(\inf K) - 1, (\sup K) + 1]$. (So K' is compact but slightly bigger than K.) By the Heine–Cantor theorem, for any $\varepsilon > 0$, there exists a $\delta_0 > 0$ such that for all $r_1, r_2 \in K'$ such that $|r_1 - r_2| < \delta_0$, we have $|\sigma(r_1) - \sigma(r_2)| < \varepsilon$. Therefore, for any $\delta \leq \min\{1, \delta_0\}$, we have

$$|\sigma_\delta(r) - \sigma(r)| \leq \int_{-\delta}^{\delta} |\sigma(r-t) - \sigma(r)| \phi_\delta(t)\, dt < \varepsilon, \qquad \forall\, r \in K.$$

Therefore,

$$\limsup_{\delta \to 0} \sup_{r \in K} |\sigma_\delta(r) - \sigma(r)| \leq \varepsilon.$$

Since this holds for all ε, we conclude the statement. □

Lemma 9. *For any $\sigma \in \mathcal{C}(\mathbb{R})$ and $\delta > 0$, $\sigma_\delta \in \mathcal{C}^\infty(\mathbb{R})$.*

Proof. With a change of variables, we can also write

$$\sigma_\delta(r) = \int_{\mathbb{R}} \sigma(t) \phi_\delta(r - t)\, dt.$$

Then

$$\begin{aligned}
\frac{d}{dr}\sigma_\delta(r) &= \lim_{h \to 0} \int_{\mathbb{R}} \sigma(t) \frac{\phi_\delta(r + h - t) - \phi_\delta(r - t)}{h}\, dt \\
&= \lim_{h \to 0} \int_{r-2\delta}^{r+2\delta} \sigma(t) \frac{\phi_\delta(r + h - t) - \phi_\delta(r - t)}{h}\, dt \\
&= \int_{r-2\delta}^{r+2\delta} \sigma(t) \lim_{h \to 0} \frac{\phi_\delta(r + h - t) - \phi_\delta(r - t)}{h}\, dt \\
&= \int_{r-2\delta}^{r+2\delta} \sigma(t) \phi_\delta'(r - t)\, dt \\
&= \int_{\mathbb{R}} \sigma(t) \phi_\delta'(r - t)\, dt
\end{aligned}$$

By Lebesgue dominated convergence theorem, since

$$\left| \frac{\phi_\delta(r + h - t) - \phi_\delta(r - t)}{h} \right| \leq \|\phi_\delta''\|_\infty.$$

By the same reasoning, we have

$$\sigma_\delta^{(k)}(r) = \int_{\mathbb{R}} \sigma(t)\phi_\delta^{(k)}(r-t)\, dt.$$

□

Lemma 10. *Let $\sigma \in \mathcal{C}(\mathbb{R})$. Then $\sigma_\delta \in \mathcal{C}^\infty(\mathbb{R})$ and $\sigma_\delta \in \overline{\operatorname{span}\{\sigma(r-t) \mid t \in \mathbb{R}\}}$, where the closure is taken in $(\mathcal{C}(K), \|\cdot\|_\infty)$ for any compact $K \subseteq \mathbb{R}$.*

Proof. Consider the Riemann sum approximation

$$\sigma_\delta(r) = \int_{\mathbb{R}} \sigma(r-t)\phi_\delta(t)\, dt \approx \frac{2\delta}{N} \sum_{i=1}^{N} \phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) \sigma\left(r + \delta - \frac{i2\delta}{N}\right).$$

To complete the proof, one must show uniform convergence

$$\lim_{N\to\infty} \sup_{r\in K} \left| \frac{2\delta}{N} \sum_{i=1}^{N} \phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) \sigma\left(r + \delta - \frac{i2\delta}{N}\right) - \sigma_\delta(r) \right| = 0.$$

As $N \to \infty$,

$$\begin{aligned}
&\left| \frac{2\delta}{N} \sum_{i=1}^{N} \phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) \sigma\left(r + \delta - \frac{i2\delta}{N}\right) - \sigma_\delta(r) \right| \\
&= \left| \sum_{i=1}^{N} \frac{2\delta}{N} \phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) \sigma\left(r + \delta - \frac{i2\delta}{N}\right) - \sum_{i=1}^{N} \int_{-\delta+\frac{(i-1)2\delta}{N}}^{-\delta+\frac{i2\delta}{N}} \phi_\delta(t)\sigma(r-t)dt \right| \\
&= \left| \sum_{i=1}^{N} \int_{-\delta+\frac{(i-1)2\delta}{N}}^{-\delta+\frac{i2\delta}{N}} \left(\phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) \sigma\left(r + \delta - \frac{i2\delta}{N}\right) - \phi_\delta(t)\sigma(r-t) \right) dt \right| \\
&\le \sum_{i=1}^{N} \int_{-\delta+\frac{(i-1)2\delta}{N}}^{-\delta+\frac{i2\delta}{N}} \left| \phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) \sigma\left(r + \delta - \frac{i2\delta}{N}\right) - \phi_\delta(t)\sigma(r-t) \right| dt \\
&\overset{\text{(i)}}{\le} \sum_{i=1}^{N} \int_{-\delta+\frac{(i-1)2\delta}{N}}^{-\delta+\frac{i2\delta}{N}} \left| \phi_\delta\left(-\delta + \frac{i2\delta}{N}\right) - \phi_\delta(t) \right| \left| \sigma\left(r + \delta - \frac{i2\delta}{N}\right) \right| \\
&\qquad\qquad + \left| \sigma\left(r + \delta - \frac{i2\delta}{N}\right) - \sigma(r-t) \right| |\phi_\delta(t)|\, dt \\
&\overset{\text{(ii)}}{\le} \sum_{i=1}^{N} \int_{-\delta+\frac{(i-1)2\delta}{N}}^{-\delta+\frac{i2\delta}{N}} \left| \varepsilon \sup_{r\in K} |\sigma(r)| + \varepsilon\frac{1}{\delta} \right| dt = 2\varepsilon(\delta \sup_{r\in K} |\sigma(r)| + 1).
\end{aligned}$$

Inequality (i) follows from the argument $|ab - cd| = |(a-c)b + c(b-d)| \leq |(a-c)b| + |c(b-d)|$. Inequality (ii) follows from the following two arguments: by the Heine–Cantor theorem, for any $\varepsilon > 0$, there exists a large enough $N \in \mathbb{N}$ such that $|\phi_\delta(x) - \phi_\delta(y)| < \varepsilon$ and $|\sigma(x) - \sigma(y)| < \varepsilon$ for any $x, y \in \overline{\mathcal{B}(K, 2\delta/N)}$ and $|x - y| < 2\delta/N$; also, $\sup_{r\in\mathbb{R}} |\phi_\delta(t)| \approx 0.83/\delta \leq 1/\delta$. Since this bound holds for any $\varepsilon > 0$, we conclude the convergence is uniform. □

Lemma 11. *Let $\sigma \in \mathcal{C}(\mathbb{R})$ be non-polynomial. Let $k \in \mathbb{N}$. There is a $\delta > 0$ such that σ_δ is not a polynomial of degree at most k.*

(σ_δ is probably not a polynomial in most cases, for any $\delta > 0$. However, we leave open the possibility that σ_δ is a polynomial of large degree; since $\sigma_\delta \to \sigma$ uniformly, there must be a sufficiently small δ such that σ_δ is either not a polynomial or a polynomial with degree larger than k.)

Proof. The set of polynomials of degree at most k is a closed subspace in $(\mathcal{C}(K), \|\cdot\|_\infty)$ for any compact $K \subset \mathbb{R}$. If σ_δ is a polynomial of degree at most k for all $\delta > 0$, then the limit $\sigma_\delta \to \sigma$ as $\delta \to 0$ must be a polynomial of degree at most k. This contradicts that assumption that σ is not polynomial, and we conclude the statement. □

Lemma 12. *Let $\sigma \in \mathcal{C}(\mathbb{R})$ be non-polynomial. Then* $\operatorname{span}\{\sigma(sr+b) \mid s \in \mathbb{R},\, t \in \mathbb{R}\}$ *is dense in any $\mathcal{C}(K)$ for any compact $K \subseteq \mathbb{R}$.*

Proof. By Lemma 10,

$$\bigcup_{\delta>0} \overline{\operatorname{span}\{\sigma_\delta(sr+t) \mid s \in \mathbb{R},\, t \in \mathbb{R}\}} \subseteq \overline{\operatorname{span}\{\sigma(sr+t) \mid s \in \mathbb{R},\, t \in \mathbb{R}\}}.$$

For any $k \in \mathbb{N}$, by Lemma 11, there exists a $\delta > 0$ such that σ_δ is not a polynomial of degree at most $k-1$, and, by Lemmas 6 and 7, r^k is in the LHS. Since the LHS and the RHS contains all monomials and therefore all polynomials, the RHS is dense by Stone–Weierstrass. □

1.3 Interpolation

Let us now address the question of interpolation: given $X_1, \ldots, X_N \in \mathbb{R}^d$ and $Y_1, \ldots, Y_N \in \mathbb{R}$, is there an θ such that $f_\theta(X_i) = Y_i$ for all $i = 1, \ldots, N$? The idea is that we have the observations $f_\star(X_i) = Y_i$ for $i = 1, \ldots, N$ of the true unknown function $f_\star$. Instead of approximating $f_\star$ on all possible inputs, the goal of interpolation is to match $f_\star$ only on the observed points.

Let $h_1, \ldots, h_N$ be functions mapping from $\mathcal{A}$ to $\mathbb{R}$. We say $\{h_i\}_{i=1}^N$ is linearly independent as functions if there does not exist a nonzero $u \in \mathbb{R}^N$ such that

$$h(a) = \sum_{i=1}^{N} u_i h_i(a) = 0, \qquad \forall\, a \in \mathcal{A}.$$

Lemma 13. *Let $h_1, \ldots, h_N$ be functions from $\mathcal{A}$ to $\mathbb{R}$. If $\{h_i\}_{i=1}^N$ are linearly independent as functions, there is $a_1, \ldots, a_N \in \mathcal{A}$ such that $M \in \mathbb{R}^{N\times N}$ defined by $M_{ij} = (h_i(a_j))$ is invertible. Furthermore, for any $Y_1, \ldots, Y_N \in \mathbb{R}$, there is a $u \in \mathbb{R}^N$ such that*

$$Y_j = \sum_{i=1}^{N} u_i h_i(a_j), \qquad \forall\, j = 1, \ldots, N,$$

i.e.,

$$\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_N \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix} M.$$

Proof. Define $H\colon \mathcal{A} \to \mathbb{R}^N$ as

$$H(a) = \begin{bmatrix} h_1(a) \\ h_2(a) \\ \vdots \\ h_N(a) \end{bmatrix}.$$

Then $\{h_i\}_{i=1}^N$ is linearly independent if and only if for $u \in \mathbb{R}^N$,

$$[u^T H(a) = 0,\ \forall\, a \in \mathcal{A}] \quad \Rightarrow \quad u = 0.$$

Also,

$$M = \begin{bmatrix} H(a_1) & H(a_2) & \cdots & H(a_N) \end{bmatrix} \in \mathbb{R}^{N\times N}$$

for any $a_1, \ldots, a_n \in \mathcal{A}$.

We establish the claim by induction. By linear independence, there is an $a_1 \in \mathcal{A}$ such that

$$v^{(1)} = H(a_1) \neq 0.$$

Next, for $i = 1, \ldots, N-1$, assume $\{v^{(1)}, \ldots, v^{(i)}\}$ is linearly independent as vectors in $\mathbb{R}^N$. Define $V^{(i)} = \operatorname{span}\{v^{(1)}, \ldots, v^{(i)}\} \subset \mathbb{R}^N$ and find a nonzero $u^{(i)} \in (V^{(i)})^{\perp}$. Then there is a $a_{i+1} \in \mathcal{A}$ such that $(u^{(i)})^{\intercal} H(a_{i+1}) \neq 0$. With $v^{(i+1)} = H(a_{i+1})$, $\{v^{(1)}, \ldots, v^{(i+1)}\}$ is linearly independent. (Since $u^{(i)}$ is orthogonal to all vectors in $V^{(i)}$, $(v^{(i+1)})^{\intercal} u^{(i)} \neq 0$ implies $v^{(i+1)} \notin V^{(i)}$.) When this process concludes at $i = N$,

$$M = \begin{bmatrix} v^{(1)} & \cdots & v^{(N)} \end{bmatrix} = \begin{bmatrix} H(a_1) & H(a_2) & \cdots & H(a_N) \end{bmatrix}$$

is invertible. □

Theorem 8 (Interpolation)**.** *Let $\sigma \in \mathcal{C}(\mathbb{R})$ be non-polynomial. Let $X_1, \dots, X_N \in \mathbb{R}^d$ be distinct data points with corresponding labels $Y_1, \dots, Y_N \in \mathbb{R}$. There exists $a_1, \dots, a_N \in \mathbb{R}^d$, $b_1, \dots, b_N \in \mathbb{R}$, and $u_1, \dots, u_N \in \mathbb{R}$ such that*

$$Y_j = \sum_{i=1}^{N} u_i \sigma(a_i^\intercal X_j + b_i), \qquad \forall\, j = 1, \dots, N.$$

Proof. Define $h_i \colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ as

$$h_i(a, b) = \sigma(a^\intercal X_i + b)$$

for $i = 1, \dots, N$. If $\{h_i\}_{i=1}^N$ are linearly independent as functions, then we are done by Lemma 13.

Now assume for contradiction that $\{h_i\}_{i=1}^N$ is linearly dependent, i.e., there is a nonzero $(u_1, \dots, u_N) \in \mathbb{R}^N$ such that

$$\sum_{i=1}^{N} u_i \sigma(a^\intercal X_i + b) = 0, \qquad \forall\, a \in \mathbb{R}^n,\, b \in \mathbb{R}.$$

If we define

$$\mu = \sum_{i=1}^{N} u_i \delta_{X_i}, \qquad L_\mu[f] = \int_\Omega f(x)\, d\mu(x)$$

then

$$L_\mu[\sigma(a^\intercal \cdot + b)] = \int_\Omega \sigma(a^\intercal x + b)\, d\mu(x) = 0, \qquad \forall\, a \in \mathbb{R}^n,\, b \in \mathbb{R}.$$

Then $L_\mu \colon \mathcal{C}(\Omega) \to \mathbb{R}$ is a bounded linear form and it vanishes on

$$\overline{\operatorname{span}(\{\sigma(a^\intercal \cdot + b) \,|\, a \in \mathbb{R}^d,\, b \in \mathbb{R}\})} = \mathcal{C}(\Omega).$$

So $L_\mu = 0$ and, by the Riesz–Markov–Kakutani representation theorem, $\mu = 0$. This is a contradiction, and we are forced to conclude that $\{h_i\}_{i=1}^N$ is linearly independent. □

1.4 Density in L^p spaces

Now that we have established density of

$$\mathcal{S}^d = \operatorname{span}\{\sigma(a^\intercal \cdot + b) \,|\, a \in \mathbb{R}^d,\, b \in \mathbb{R}^n\}$$

in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$ for any compact $\Omega \subseteq \mathbb{R}^d$, one may wonder whether $\mathcal{S}^d$ is dense in L^p spaces.

For $p \in [1, \infty)$, the usual L^p space with respect to the Lebesgue measure is defined as the vector space of (equivalence classes of) functions f such that

$$\|f\|_{L^p}^p = \int_{\mathbb{R}^d} |f(x)|^p \, dx < \infty.$$

However, $\mathcal{S}^d$ cannot be dense in L^p.

Theorem 9. *[Chui, Li, Mhaskar (1994)?] Let $d \geq 2$. For any (Lebesgue measurable) $\sigma \colon \mathbb{R} \to \mathbb{R}$, any nonzero $g \in \mathcal{S}^d$ satisfies*

$$\|g\|_{L^p} = \infty$$

for all $p \in [1, \infty)$.

The proof of this result seems somewhat tricky, so we omit it. The issue is that 2-layer neural networks cannot effectively approximate (in $\|\cdot\|_\infty$ or $\|\cdot\|_{L^p}$) a function compact support. We return to this point when we discuss 3-layer neural networks.

However, $\mathcal{S}^d$ is dense in $L^p(\mu)$ for $p \in [1, \infty)$. Let $\mu \in \mathcal{M}_+(\mathbb{R}^d)$ be a finite nonnegative measure. For $p \in [1, \infty)$, the $L^p(\mu)$ space is defined as the vector space of (equivalence classes of) functions f such that

$$\|f\|_{L^p(\mu)}^p = \int_{\mathbb{R}^d} |f(x)|^p \, d\mu(x) < \infty.$$

(Note the Lebesgue measure is not a finite measure.)

Finally, we point out that $\mathcal{S}^d$ cannot be dense in $L^\infty(\mu)$ since the continuous functions on $\mathcal{S}^d$ cannot approximate discontinuous functions in the $\|\cdot\|_{L^\infty(\mu)}$-norm.

Theorem 10. *Let $p \in [1, \infty)$. Let $\mu \in \mathcal{M}_+(\mathbb{R}^d)$. Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying*

$$\lim_{r \to -\infty} \sigma(r) = 0, \qquad \lim_{r \to \infty} \sigma(r) = 1.$$

Then $\mathcal{S}^d$ is dense in $L^p(\mu)$.

Proof. Since μ is finite, $\mathcal{S}^d \subset L^p(\mu)$. Assume for contradiction that $\mathcal{S}^d$ is not dense in $L^p(\mu)$. We use the Hahn–Banach extension theorem as in Lemma 1 to obtain a nonzero bounded linear functional $L \colon L^p(\mu) \to \mathbb{R}$ such that

$$L[f] = 0, \qquad \forall\, f \in \overline{\mathcal{S}^d}.$$

Let $q = p/(p-1)$, with $q = \infty$ for $p = 1$. Since $(L^p(\mu))^* = L^q(\mu)$, there is a $g \in L^q(\mu)$ such that

$$L[f] = \int_{\mathbb{R}^d} fg \, d\mu, \qquad \forall \, f \in L^p(\mu).$$

Let $d\nu = g d\mu$, i.e.,

$$\nu(A) = \int_A g \, d\mu, \qquad \forall \text{ measurable } A \subseteq \mathbb{R}^d.$$

By Hölder's inequality, $\|g\|_{L^1(\mu)} \leq \mu(\mathbb{R}^d)\|g\|_{L^q(\mu)} < \infty$. Therefore $g \in L^1(\mu)$, and ν is a finite signed measure, i.e., $\nu \in \mathcal{M}(\mathbb{R}^d)$. Then

$$\int_{\mathbb{R}^d} \sigma(a^\intercal x + b) \, d\nu(x) = L[\sigma(a^\intercal \cdot + b)] = 0, \qquad \forall \, a \in \mathbb{R}^d, \, b \in \mathbb{R}.$$

However, σ is discriminatory by Lemma 2, so $\nu = 0$. This contradicts the construction of L as a nonzero linear form. Therefore, we are forced to conclude that $\mathcal{S}^d$ is dense in $L^p(\mu)$. □

1.5 Quantitative approximation guarantees by probabilistic method

Our prior results on approximation capabilities are existence results; they come with no quantitative bounds on how large N must be to attain an ε-approximation. Let us now consider a probabilistic construction (still not a practical construction) to obtain quantitative results.

The probabilistic method is a proof technique pioneered by Paul Erdös. We separately illustrate the technique with the following example.

Fact 1. 10% *of the surface of a sphere is colored blue, the rest is red. Show that, there is an inscribed cube with all its vertices touching red.*

Proof. Let B_r be the event that the r-th vertex of a randomly selected cube is touches blue and note that $\mathbb{P}(B_r) = 1/10$. By an application of the union bound,

$$\mathbb{P}\left[\text{At least one corner touch blue}\right] = \mathbb{P}\left[\bigcup_{r=1}^{8} B_r\right] \leq \sum_{r=1}^{8} \mathbb{P}(B_r) = \frac{8}{10} < 1$$

and therefore

$$\mathbb{P}[\text{All corners touch red}] = \mathbb{P}\left[\left(\bigcup_{r=1}^{8} B_r\right)^{\complement}\right] = 1 - \mathbb{P}\left[\bigcup_{r=1}^{8} B_r\right] > 0.$$

Since there is positive probability all vertices touching red, there must exist a (non-random) configuration with all vertices touching red. □

For $B \in (0, \infty)$, define the $L^2(B)$-norm $\|\cdot\|_{L^2(B)}$ as

$$\|f\|_{L^2(B)}^2 = \int_{\mathcal{B}(0,B)} (f(x))^2 \, dx.$$

where $\mathcal{B}(0, B)$ is the closed ball of radius B centered at 0. (So $L^2(B) = L^2(\mu)$ where μ is defined by $d\mu = \mathbf{1}_{\mathcal{B}(0,B)} dx$.)

We approximate a given $f_\star$ via the probabilistic method with the following outline. First, find a $\tilde{f}_\star \approx f_\star$ such that

$$\tilde{f}_\star(x) = \mathbb{E}_{(w,b)\sim P}[c(w,b)\sigma(w^\intercal x + b)] = \int_{\mathbb{R}^d \times \mathbb{R}} c(w,b)\sigma(w^\intercal x + b) \, dP(w,b)$$

for some $c(w, b) \leq C < \infty$ and probability measure P on $\mathbb{R}^d \times \mathbb{R}$. Then, sample $(w_1, b_1), \ldots, (w_N, b_N) \sim P$ i.i.d. and form

$$f_\theta(x) = \sum_{i=1}^{N} \frac{c(w_i, b_i)}{N} \sigma(w_i^\intercal x + b_i).$$

Since $\mathbb{E}[f_\theta] = \tilde{f}_\star$, and

$$\mathbb{E}_\theta \|f_\theta - \tilde{f}_\star\|_{L^2(B)}^2 \leq \frac{\text{Variance}}{N},$$

there exists a θ such that

$$\|f_\theta - \tilde{f}_\star\|_{L^2(B)}^2 \leq \frac{\text{Variance}}{N}.$$

(The variance will be finite under the given assumptions.)

Theorem 11. *Let $B \in (0, \infty)$. Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying*

$$\lim_{r \to -\infty} \sigma(r) = 0, \qquad \lim_{r \to \infty} \sigma(r) = 1, \qquad |\sigma(r)| \leq 1, \, \forall \, r \in \mathbb{R}.$$

Let $f_\star \colon \mathbb{R}^d \to \mathbb{R}$ be a function with an absolutely integrable Fourier representation $\hat{f}_\star \colon \mathbb{R}^d \to \mathbb{C}$, i.e.,

$$f_\star(x) = \int_{\mathbb{R}^d} e^{-iw^\intercal x} \hat{f}_\star(w)\, dw,\ \forall\, x \in \mathbb{R}^d, \qquad \int_{\mathbb{R}^d} |\hat{f}_\star(w)|\, dw < \infty.$$

Further assume that

$$Q = \int_{\mathbb{R}^d} \|w\| |\hat{f}_\star(w)|\, dw < \infty.$$

Then for any $N \in \mathbb{N}$, there exists

$$f_\theta(x) = \sum_{i=1}^{N+1} \lambda_i \sigma(w_i^\intercal x + b_i)$$

such that

$$\|f_\theta - f_\star\|^2_{L^2(B)} \leq \frac{5Q^2B^2 \operatorname{Vol}(\mathcal{B}(0,B))}{N}.$$

As an aside, the assumption $Q < \infty$ is a sufficient condition ensuring the gradient exists and can that it can be evaluated under the Fourier integral (by Lebesgue dominated convergence theorem):

$$\begin{aligned}
\nabla f_\star(x) &= \nabla \int_{\mathbb{R}^d} e^{-iw^\intercal x} \hat{f}_\star(w)\, dw \\
&= \int_{\mathbb{R}^d} \nabla e^{-iw^\intercal x} \hat{f}_\star(w)\, dw \\
&= \int_{\mathbb{R}^d} -iw e^{-iw^\intercal x} \hat{f}_\star(w)\, dw.
\end{aligned}$$

We establish Theorem 11 via the following lemmas.

Lemma 14 (Maurey, Pilsner, Jones [3, 1])**.** *Let $\mathcal{H}$ be a Hilbert space. Let $(\mathcal{W}, P)$ be a probability space.*[4] *Let $h \colon \mathcal{W} \to \mathcal{H}$ such that $\|h(w)\| \leq H < \infty$ for (P-almost) all $w \in \mathcal{W}$. Assume*

$$f = \int_{\mathcal{W}} h(w)\, dP(w) = \mathbb{E}_{w \in P}[h(w)].$$

Then, for any $N \in \mathbb{N}$, there exists $h_1, \ldots, h_N \in \mathcal{H}$, such that

$$\tilde{f} = \sum_{i=1}^{N} \frac{1}{N} h_i$$

[4]We use $\mathcal{W}$ to represent the sample space, rather than the more common Ω, since Ω denotes a compact subset of $\mathbb{R}^d$ for us. We omit specifying the σ-algebra.

satisfies

$$\|\tilde{f} - f\|^2 \leq \frac{H^2}{N}.$$

Proof. Sample $w_1, \ldots, w_N \sim P$ i.i.d. Then

$$\hat{f} = \sum_{i=1}^{N} \frac{1}{N} h(w_i)$$

has

$$\mathbb{E}[\hat{f}] = f$$

and

$$\mathbb{E}[\|\hat{f} - f\|^2] = \frac{1}{N}\mathbb{E}[\|h(w_1) - f\|^2] = \frac{1}{N}(\mathbb{E}[\|h(w_1)\|^2] - \|f\|^2) \leq \frac{H^2 - \|f\|^2}{N} \leq \frac{H^2}{N}.$$

Since $\hat{f}$ is a random variable with variance at most H^2/N, there is a particular (non-random) instance $\tilde{f}$ such that $\|\tilde{f} - f\|^2 \leq H^2/N$. □

Lemma 15. *Let B and $f_\star$ satisfy the assumptions of Theorem 11. Then there exists $\varphi \colon \mathbb{R}^d \to [0, 2\pi)$ and a probability measure P on $\mathbb{R}^d \times \mathbb{R}$ such that it is absolutely continuous with respect to the Lebesgue measure and*

$$f_\star(x) - f_\star(0) = 2BQ \int_{\mathbb{R}^d \times \mathbb{R}} \sin(b - \varphi(w)) \mathbf{1}_{\{w^\intercal x + b \geq 0\}} \, dP(w, b)$$

for all $x \in \mathcal{B}(0, B)$.

Proof. Define $\varphi(w) \in [0, 2\pi)$ such that $\hat{f}_\star(w) = e^{-i\varphi(w)} |\hat{f}_\star(w)|$. Then,

$$\begin{aligned} f_\star(x) &= \int_{\mathbb{R}^d} e^{-iw^\intercal x - i\varphi(w)} |\hat{f}_\star(w)| \, dw \\ &= \Re \int_{\mathbb{R}^d} e^{-iw^\intercal x - i\varphi(w)} |\hat{f}_\star(w)| \, dw \\ &= \int_{\mathbb{R}^d} \cos(w^\intercal x + \varphi(w)) |\hat{f}_\star(w)| \, dw, \end{aligned}$$

and

$$f_\star(x) - f_\star(0) = \int_{\mathbb{R}^d} (\cos(w^\intercal x + \varphi(w)) - \cos(\varphi(w))) |\hat{f}_\star(w)| \, dw.$$

Next, we have

$$\begin{aligned}
&\cos(w^\intercal x + \varphi(w)) - \cos(\varphi(w)) \\
&= -\int_0^{w^\intercal x} \sin(b + \varphi(w))\, db \\
&= -\int_0^{B\|w\|} \mathbf{1}_{\{w^\intercal x - b \geq 0\}} \sin(b + \varphi(w))\, db + \int_{-B\|w\|}^{0} \mathbf{1}_{\{-w^\intercal x + b \geq 0\}} \sin(b + \varphi(w))\, db,
\end{aligned}$$

where the two terms correspond to the cases where $w^\intercal x \geq 0$ and $w^\intercal x < 0$. We also use $w^\intercal x \leq \|w\|\|x\| \leq B\|w\|$ to restrict the integration boundaries. Then we have

$$\begin{aligned}
f_\star(x) - f_\star(0) &= -\int_{\mathbb{R}^d}\int_{\mathbb{R}} \mathbf{1}_{\{w^\intercal x - b \geq 0\}} \sin(b + \varphi(w))\, \mathbf{1}_{\{0 \leq b \leq B\|w\|\}} |\hat{f}_\star(w)|\, dbdw \\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}} \mathbf{1}_{\{-w^\intercal x + b \geq 0\}} \sin(b + \varphi(w))\, \mathbf{1}_{\{-B\|w\| \leq b \leq 0\}} |\hat{f}_\star(w)|\, dbdw \\
&= -\int_{\mathbb{R}^d}\int_{\mathbb{R}} \mathbf{1}_{\{w^\intercal x + b \geq 0\}} \sin(-b + \varphi(w))\, \mathbf{1}_{\{-B\|w\| \leq b \leq 0\}} |\hat{f}_\star(w)|\, dbdw \\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}} \mathbf{1}_{\{w^\intercal x + b \geq 0\}} \sin(b - \varphi(w))\, \mathbf{1}_{\{-B\|w\| \leq b \leq 0\}} |\hat{f}_\star(w)|\, dbdw \\
&= 2\int_{\mathbb{R}^d}\int_{\mathbb{R}} \mathbf{1}_{\{w^\intercal x + b \geq 0\}} \sin(b - \varphi(w))\, \mathbf{1}_{\{-B\|w\| \leq b \leq 0\}} |\hat{f}_\star(w)|\, dbdw \\
&= 2BQ \int_{\mathbb{R}^d \times \mathbb{R}} \sin(b - \varphi(w)) \mathbf{1}_{\{w^\intercal x + b \geq 0\}}\, dP(w, b)
\end{aligned}$$

where we used the property that $\hat{f}_\star(w) = \overline{\hat{f}_\star}(-w)$ since $f_\star$ is real. For the final step, define $dP \propto \mathbf{1}_{\{-B\|w\| \leq b \leq 0\}} |\hat{f}_\star(w)|\, dbdw$ with the normalization factor

$$2\int_{\mathbb{R}^d}\int_{\mathbb{R}} \mathbf{1}_{\{-B\|w\| \leq b \leq 0\}} |\hat{f}_\star(w)| dbdw = 2B \int_{\mathbb{R}^d} \|w\| |\hat{f}_\star(w)| dw = 2BQ.$$

□

Lemma 16. *Let σ satisfy the assumptions of Theorem 11. Let $|s(w, b)| \leq 1$ for all w and b. Let*

$$h(x) = \int_{\mathbb{R}^d \times \mathbb{R}} s(w, b) \mathbf{1}_{\{w^\intercal x + b \geq 0\}}\, dP(w, b),$$

where P is a probability measure that is absolutely continuous with respect to the Lebesgue measure. Then for any $\delta > 0$, there are s^δ, such that $|s^\delta(w, b)| \leq 1$ for all w and b, and a probability measure P^δ such that

$$h_\delta(x) = \int_{\mathbb{R}^{d+1}} s^\delta(w, b) \sigma(w^\intercal x + b)\, dP^\delta(w, b)$$

satisfies

$$\|h_\delta - h_\star\|_{L^2(B)} \to 0$$

as $\delta \to 0$.

Proof. By the Lebesgue dominated convergence theorem, we have

$$\int_{\mathbb{R}^{d+1}} (s(w,b))^2 \left(\sigma\left(\frac{w^\intercal x}{\delta} + \frac{b}{\delta}\right) - \mathbf{1}_{\{w^\intercal x + b \geq 0\}}\right)^2 \, dP(w,b) \to 0$$

as $\delta \to 0$. Finally, we use the change of variables $\tilde{w} = w/\delta$ and $\tilde{b} = b/\delta$, let $s^\delta(\tilde{w}, \tilde{b}) = s(\delta\tilde{w}, \delta\tilde{b})$, and let P^δ be the probability measure on $\tilde{w}$ and $\tilde{b}$. Then

$$\int_{\mathbb{R}^{d+1}} s(w,b)\sigma\left(\frac{w^\intercal}{\delta}x + \frac{w}{\delta}\right) \, dP(w,b) = \int_{\mathbb{R}^{d+1}} s^\delta(\tilde{w},\tilde{b})\sigma\left(\tilde{w}^\intercal x + \tilde{b}\right) \, dP^\delta(\tilde{w},\tilde{b}).$$

□

Proof of Theorem 11. By Lemma 15, we can find

$$f_\star(x) - f_\star(0) = 2BQ \int_{\mathbb{R}^{d+1}} s(w,b)\mathbf{1}_{\{w^\intercal x + b \geq 0\}} \, dP(w,b)$$

such that $|s(w,b)| \leq 1$ By Lemma 16, we can find

$$\widetilde{\Delta f}_\star(x) = 2BQ \int_{\mathbb{R}^{d+1}} s^\delta(w,b)\sigma(w^\intercal x + b) \, dP^\delta(w,b)$$

such that $|s^\delta(w,b)| \leq 1$ and

$$\|\widetilde{\Delta f}_\star(\cdot) - f_\star(\cdot) + f_\star(0)\|^2 \leq \frac{(\sqrt{5}-2)^2 B^2 Q^2 \operatorname{Vol}(\mathcal{B}(0,B))}{N}.$$

Then by Lemma 14, we there exists a

$$f_{\theta'}(x) = \sum_{i=1}^N \lambda_i \sigma(w_i^\intercal x + b_i)$$

such that

$$\|f_{\theta'} - \widetilde{\Delta f}_\star\|^2 \leq \frac{4B^2Q^2 \operatorname{Vol}(\mathcal{B}(0,B))}{N}.$$

Finally, we let $\lambda_{N+1} = f_\star(0)/\sigma(b_{N+1})$, $w_{N+1} = 0$, and $b_{N+1} \in \mathbb{R}$ be such that $\sigma(b_{N+1}) \neq 0$, then

$$f_\theta(x) = \sum_{i=1}^{N+1} \lambda_i \sigma(w_i^\intercal x + b_i)$$

satisfies, by the triangle inequality,

$$\|f_\theta - f_\star\|_{L^2(B)}^2 \leq \frac{5B^2Q^2 \operatorname{Vol}(\mathcal{B}(0,B))}{N}.$$

□

1.6 Approximation capabilities of deeper neural networks

Consider 3-layer neural networks of the form

$$f_\theta(x) = A^{(3)}\sigma_2(A^{(2)}\sigma_1(A^{(1)}x + b^{(1)}) + b^{(2)}) + b^{(3)}, \tag{1.3}$$

where $N_1, N_2 \in \mathbb{N}$, $A^{(1)} \in \mathbb{R}^{N_1 \times d}$, $b^{(1)} \in \mathbb{R}^{N_1}$, $A^{(2)} \in \mathbb{R}^{N_2 \times N_1}$, $b^{(2)} \in \mathbb{R}^{N_2}$, $A^{(3)} \in \mathbb{R}^{1 \times N_2}$, $b^{(3)} \in \mathbb{R}^1$, $\sigma_1 \colon \mathbb{R} \to \mathbb{R}$, $\sigma_2 \colon \mathbb{R} \to \mathbb{R}$, and σ_1 and σ_2 are applied elementwise.

Since 2-layer neural networks are already universal approximators, why consider 3-layer or deeper neural networks? The empirical observation is clear: deeper neural networks perform far better than 2-layer neural networks.

Although our theoretical understanding of the effectiveness of depth is far from complete, there are some known results on their approximation capabilities. In this section, we quickly introduce some, mostly without providing complete proofs.

1.6.1 Approximating compactly supported functions

As discussed in Theorem 9, 2-layer neural networks cannot effectively approximate compactly supported functions on all of $\mathbb{R}^d$. However, 3-layer neural networks can.

To understand why, consider the following example. Let

$$A = \begin{bmatrix} a_1^\intercal \\ a_2^\intercal \\ \vdots \\ a_m^\intercal \end{bmatrix} \in \mathbb{R}^{m \times d}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m.$$

Consider the indicator function on the convex polytype

$$\mathbf{1}_{\{x \,|\, Ax \le b\}}(x) = \begin{cases} 1 & a_i^\intercal x \le b_i, \text{ for all } i = 1, \dots, m \\ 0 & \text{otherwise,} \end{cases}$$

where the inequality in $Ax \le b$ is elementwise. Let

$$s(r) = \begin{cases} 0 & \text{for } r < 0 \\ 1 & \text{for } r \ge 0. \end{cases}$$

be the step function. Then

$$s\left(\sum_{i=1}^m s(b_i - a_i^\intercal \cdot) - m + \frac{1}{2}\right) = \mathbf{1}_{\{x \,|\, Ax \le b\}}.$$

If $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a continuous function satisfying

$$\lim_{r\to-\infty} \sigma(r) = 0, \qquad \lim_{r\to\infty} \sigma(r) = 1,$$

then

$$\sigma\left(\frac{1}{\delta}\left(\sum_{i=1}^{m} \sigma\left(\frac{1}{\delta}(b_i - a_i^\intercal \cdot)\right) - m + \frac{1}{2}\right)\right) \to \mathbf{1}_{\{x \,|\, Ax \le b\}}(x)$$

as $\delta \to 0$ for almost all x.

1.6.2 Universality of 3-layer wide neural networks

Universality of wide neural networks of depth 3 or deeper, requires a little bit of additional work to establish.

Theorem 12. *Let $\Omega \subseteq \mathbb{R}^d$ be compact. Assume $\sigma_2 \in \mathcal{C}^\infty(\mathbb{R})$ and $\sigma_1 \in \mathcal{C}(\mathbb{R})$ are non-polynomial. Then the class of 3-layer neural networks of the form (1.3) are dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.*

Proof. Since σ_2 is non-polynomial, there is a point r_0 such that $\sigma_2'(r_0) \neq 0$. For any 2-layer neural network $h \in \operatorname{span}(\{\sigma_1(a^\intercal \cdot + b) \,|\, a \in \mathbb{R}^d,\, b \in \mathbb{R}\})$, we have

$$f_\theta(x) = \frac{1}{\varepsilon\sigma_2'(r_0)}(\sigma_2(\varepsilon h(x) + r_0) - \sigma_2(r_0)) \to h(x)$$

uniformly for $x \in \Omega$ as $\varepsilon \to 0$. Therefore,

$$\operatorname{span}(\{\sigma_1(a^\intercal \cdot + b) \,|\, a \in \mathbb{R}^d,\, b \in \mathbb{R}\}) \subseteq \overline{\{\text{functions of the form (1.3)}\}}.$$

The left-hand side is dense by Theorem 7, so is the right-hand side.

Finally, it remains to establish the claimed uniform convergence. Let $\Omega' = \{\varepsilon h(x) + r_0 \,|\, x \in \Omega, \varepsilon \in [-1, 1]\}$. Since

$$\begin{aligned}\frac{\sigma_2(\varepsilon h(x) + r_0) - \sigma_2(r_0)}{\varepsilon\sigma_2'(r_0)} - h(x) &= \frac{h(x)}{\varepsilon\sigma_2'(r_0)} \int_0^\varepsilon (\sigma_2'(\eta h(x) + r_0) - \sigma_2'(r_0))\, d\eta \\ &= \frac{(h(x))^2}{\varepsilon\sigma_2'(r_0)} \int_0^\varepsilon \int_0^\eta \sigma_2''(\nu h(x) + r_0)\, d\nu\, d\eta,\end{aligned}$$

for $|\varepsilon| \in (0, 1)$, we have

$$\begin{aligned}&\sup_{x\in\Omega} \left|\frac{\sigma_2(\varepsilon h(x) + r_0) - \sigma_2(r_0)}{\varepsilon\sigma_2'(r_0)} - h(x)\right| \\ &\le \frac{\varepsilon}{\sigma_2'(r_0)} \left(\sup_{x\in\Omega}(h(x))^2\right) \left(\sup_{r\in\Omega'} |\sigma_2''(r)|\right) < \infty.\end{aligned}$$

□

1.6.3 Depth separation

Depth separation results establish that certain tasks cannot be done by shallower networks while deeper networks can. Since L-layer neural networks are universal approximators for $L \geq 2$, depth separation results often focus on quantitative approximation capabilities.

Consider the target function $f_\star = \mathbf{1}_{\mathcal{B}(0,1)}$. One can approximate $\mathbf{1}_{\mathcal{B}(0,1)}$ with a 3-layer neural network, by approximating $\|x\|^2 = x_1^2 + \cdots + x_d^2$ with 2-layers and then approximating $\mathbf{1}_{\{r \,|\, r \leq 1\}}$ with the third layer.

Theorem 13 (Safran and Shamir [4], Informal)**.** *Assume $\sigma \colon \mathbb{R} \to \mathbb{R}$ satisfy some conditions. For any $\mu \in \mathcal{M}_+(\mathbb{R}^d)$ and $\varepsilon > 0$, there exists a 3-layer neural network of the form* (1.3) *satisfying*

$$\|f_\theta - \mathbf{1}_{\mathcal{B}(0,1)}\|_{L^2(\mu)}^2 < \varepsilon$$

with width $\max\{N_1, N_2\} \leq \mathcal{O}(d^2/\varepsilon)$ as $d \to \infty$.

Surprisingly, however, approximating this rather simple function via a 2-layer neural network requires an inordinate width.

Theorem 14 (Safran and Shamir [4], Informal)**.** *Assume $\sigma \colon \mathbb{R} \to \mathbb{R}$ satisfy some conditions. There exists $\mu \in \mathcal{M}_+(\mathbb{R}^d)$ such that any 2-layer neural networks of the form* (1.1) *satisfying*

$$\|f_\theta - \mathbf{1}_{\mathcal{B}(0,1)}\|_{L^2(\mu)}^2 < \mathcal{O}(1/d^4)$$

must have width at least $N \geq \Omega(\exp(Cd))$ as $d \to \infty$, where C is a constant only depending on σ.

Let $d = 1$ and

$$\Delta(x) = \begin{cases} 2x & \text{for } x \in [0, 1/2) \\ 2 - 2x & \text{for } x \in [1/2, 1) \\ 0 & \text{otherwise.} \end{cases}$$

Define

$$\Delta^k = \underbrace{\Delta \circ \Delta \circ \cdots \circ \Delta}_{k \text{ times}}.$$

Then Δ^k exhibits a fractal-like behavior; Δ^k has 2^{k-1} triangular peeks of height 1 and width $1/2^{k-1}$ in $[0, 1]$. Roughly speaking, these exponentially many ups and downs can only be created through depth.

Theorem 15 (Telgarsky [5])**.** *Let $L \geq 2$. Then Δ^{L^2+2} is a ReLU network with $3L^2 + 6$ nodes and $2L^2 + 4$ layers. However, any ReLU network f_θ with at most 2^L nodes and L layers cannot approximate it:*

$$\int_0^1 |\Delta^{L^2+2}(x) - f_\theta(x)| \, dx \geq \frac{1}{32}.$$

Here, "node" refers to the sum of the widths of all layers.

Chapter 2

Positive definite kernels

> "In mathematics, a kernel is an object to which the author assigns the name K." — Jan 6, 2022, Sam Power (`@sp_monte_carlo`)[1]

Let $\mathcal{X}$ be a nonempty set. Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We say K is symmetric if $K(x, x') = K(x', x)$ for all $x, x' \in \mathcal{X}$. Given $x_1, \ldots, x_N \in \mathcal{X}$, let $G \in \mathbb{R}^{N \times N}$ be

$$G_{ij} = K(x_i, x_j), \qquad i, j \in \{1, \ldots, N\}.$$

We call G the *kernel matrix* or the *Gramian matrix* of K. Then K is a *positive definite kernel* (PDK) if G is symmetric positive semidefinite for any $N \in \mathbb{N}$ and $x_1, \ldots, x_N \in \mathcal{X}$. Equivalently, K is positive definite if it is symmetric and

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j) \geq 0$$

for all $N \in \mathbb{N}, x_1, \ldots, x_N \in \mathcal{X}$ and $c \in \mathbb{R}^N$.

The inconsistent naming warrants some clarification. A matrix $G \in \mathbb{R}^{N \times N}$ is symmetric positive definite if all eigenvalues are strictly positive ($>$) and symmetric positive **semi**definite if all eigenvalues are nonnegative ($\geq$). In contrast, a **strictly** positive definite kernel, as defined below, refers to the strict notion ($>$) while positive definite kernels correspond to the non-strict notion ($\geq$).

We say $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a strictly positive definite kernel if for any $N \in \mathbb{N}$ and distinct $x_1, \ldots, x_N \in \mathcal{X}$, the corresponding Gramian matrix G is symmetric (strictly) positive definite. Equivalently, K is strictly positive

[1] `https://twitter.com/sp_monte_carlo/status/1478783658714673159`

definite if it is symmetric and

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) > 0$$

for all $N \in \mathbb{N}, x_1, \ldots, x_N \in \mathcal{X}$ and nonzero $c \in \mathbb{R}^N$.

PDKs arise in several separate instances within machine learning. Confusingly, "kernel" is perhaps the most inconsistently overused word in mathematics. Entirely unrelated uses include kernel as the input that produces the 0 as the output, nonnegative kernels (used in the Nadaraya–Watson estimator), convolutional kernels, kernels of operating systems facilitating interactions between hardware and software components, and GPU compute kernels containing code to be executed on GPUs.

2.1 Building blocks of kernels

We now discuss the building blocks of PDKs. This machinery will allow us to construct PDKs and to identify PDKs.

2.1.1 Inner products of feature maps

Let $\phi\colon \mathcal{X} \to \mathcal{H}$ for some Hilbert space $\mathcal{H}$ (not necessarily an RKHS) equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $\| \cdot \|_{\mathcal{H}}$. We call ϕ a *feature map* for reasons that we discuss soon. Then, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

is a PDK, since, for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and $c \in \mathbb{R}^N$,

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{N} c_i \phi(x_i), \sum_{j=1}^{N} c_j \phi(x_j) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{N} c_i \phi(x_i) \right\|_{\mathcal{H}}^2 \\
&\geq 0.
\end{aligned}$$

Example: Linear kernel. The simplest instance is $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \mathbb{R}^d$, $\phi(x) = x$, and

$$K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}.$$

Example: Tensor product. Let $f_1, \ldots, f_P$ be functions from $\mathcal{X}$ to $\mathbb{R}$. Then, $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x, x') = \sum_{i=1}^{P} f_i(x) f_i(x')$$

is a PDK. Using the notation of tensor products, which we further discuss later, we can equivalently write

$$K = \sum_{i=1}^{P} f_i \otimes f_i.$$

This is analogous to expressing a matrix as a sum of P rank-1 outer products. The sum of P tensor products is actually an instance of a PDK defined through the feature map

$$\phi(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_P(x) \end{bmatrix} \in \mathbb{R}^P.$$

Example: Min kernel. Let $\mathcal{X} = [0, \infty)$. Then, $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x, x') = \min(x, x')$$

is a PDK. To see why, for $L^2(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \,|\, (\int |f(x)|^2 dx)^{1/2} < \infty\}$, let $\phi \colon \mathcal{X} \to L^2(\mathbb{R})$ be defined by $\phi(x) = \mathbf{1}_{[0,x]}$. Then

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{L^2(\mathbb{R})} = \langle \mathbf{1}_{[0,x]}, \mathbf{1}_{[0,x']} \rangle_{L^2(\mathbb{R})} = \min(x, x').$$

2.1.2 Operations preserving PDKs

Given simple PDKs, we can construct more complex PDKs through operations preserving positive definiteness: nonnegative scalings, sums, products, limits, and integrals with respect to nonnegative measures. Let K_1 and K_2 be PDKs mapping $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Then

- αK_1 for any $\alpha \geq 0$,

- $K_1 + K_2$, and
- $K_1 K_2$

are PDKs. The first two claims are clear. The third claim means $K_3 \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$K_3(x, x') = K_1(x, x') K_2(x, x'), \qquad \forall\, x, x \in \mathcal{X}$$

is PDK, and it follows from the Schur product theorem.

Theorem 16 (Schur product theorem)**.** *Let $A \in \mathbb{R}^{N \times N}$ and $B \in \mathbb{R}^{N \times N}$ be symmetric positive semidefinite. Then the Hadamard product $C = A \odot B$, defined by $C_{ij} = A_{ij} B_{ij}$ for $i, j \in \{1, \ldots, N\}$, is symmetric positive semidefinite.*

Proof. Let

$$A = \sum_{i=1}^{N} \lambda_i u_i u_i^\intercal, \qquad B = \sum_{i=1}^{N} \nu_i v_i v_i^\intercal$$

be the eigenvalue decompositions of A and B with respective orthonormal eigenvectors $u_1, \ldots, u_N$ and $v_1, \ldots, v_N$. Since $\odot$ is bilinear,

$$\begin{aligned} C &= A \odot B \\ &= \left(\sum_{i=1}^{N} \lambda_i u_i u_i^\intercal \right) \odot \left(\sum_{j=1}^{N} \nu_j v_j v_j^\intercal \right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \nu_j \left(u_i u_i^\intercal \right) \odot \left(v_j v_j^\intercal \right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \nu_j (u_i \odot v_j)(u_i \odot v_j)^\intercal \end{aligned}$$

is a sum of N^2 (rank-0 or rank-1) symmetric positive semidefinite matrices and therefore is symmetric positive semidefinite. □

Let $\{K_i\}_{i \in \mathbb{N}}$ be a sequence of PDKs mapping $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. If

$$K_\infty(x, x') = \sum_{i=1}^{\infty} K_i(x, x')$$

exists for all $x, x' \in \mathcal{X}$, then K_∞ is a PDK. Let $\{K_w\}_{w \in \mathcal{W}}$ be a family of PDKs mapping $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Let μ be a nonnegative measure on $\mathcal{W}$. If

$$K(x, x') = \int_{\mathcal{W}} K_w(x, x')\, d\mu(w)$$

is well-defined (measurable and integrable) for all $x, x' \in \mathcal{X}$, then K is a PDK.

Example: Polynomial kernel. Let $\mathcal{X} = \mathbb{R}^d$ and $p \in \mathbb{N}$. Then, $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as

$$K(x, x') = (\langle x, x' \rangle + 1)^p$$

is a PDK.

Example: Exponential kernel. Let $\mathcal{X} = \mathbb{R}^d$. Then, $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as

$$K(x, x') = \exp(\langle x, x' \rangle) = \sum_{p=0}^{\infty} \frac{1}{p!} (\langle x, x' \rangle)^p.$$

is a PDK.

Example: Cosine kernel. Let $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x, x') = \cos(x - x') = \cos(x)\cos(x') + \sin(x)\sin(x')$$

is a PDK.

Example: Kernels with integers. Let $\mathcal{X} = \mathbb{N}$. Then, $K\colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ defined as

$$K(x, x') = 2^{xx'} = e^{(\log 2)xx'}$$

is PDK.

2.1.3 Shift invariant kernels and Bochner's theorem

Let $\mathcal{X} = \mathbb{R}^d$. We say $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is *shift-invariant* if there exists a function $\kappa\colon \mathbb{R}^d \to \mathbb{R}$ such that

$$K(x, x') = \kappa(x - x').$$

Theorem 17 (Bochner)**.** *A shift-invariant $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $K(x, x') = \kappa(x - x')$ is a PDK if and only if*

$$\kappa(t) = \int_{\mathbb{R}^d} e^{-i\omega^\intercal t} d\mu(\omega)$$

for some (real) nonnegative finite measure $\mu \in \mathcal{M}_+(\mathbb{R}^d)$.

Proof of (⇐).

$$\begin{aligned} K(x,x') &= \int_{\mathbb{R}^d} e^{-i\omega^\intercal(x-x')}\, d\mu(\omega) \\ &= \Re \int_{\mathbb{R}^d} e^{-i\omega^\intercal(x-x')}\, d\mu(\omega) \\ &= \int_{\mathbb{R}^d} \cos(\omega^\intercal(x-x'))\, d\mu(\omega) \\ &= \int_{\mathbb{R}^d} (\cos(\omega^\intercal x)\cos(\omega^\intercal x') + \sin(\omega^\intercal x)\sin(\omega^\intercal x'))\; d\mu(\omega). \end{aligned}$$

We omit the proof of (⇒) which requires more work. □

Example: Sinc kernel. Let $B > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ defined as

$$K(x,x') = 2B\mathrm{sinc}(B(x-x')) = \begin{cases} \frac{2\sin(B(x-x'))}{(x-x')} & \text{if } x \neq x' \\ 0 & \text{if } x = x' \end{cases}$$

is a PDK, since

$$2B\mathrm{sinc}(B(t)) = \int_{\mathbb{R}} e^{-i\omega t}\mathbf{1}_{[-B,B]}(\omega)\, d\omega.$$

Example: Gaussian kernel. Let $\sigma > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ defined as

$$K(x,x') = e^{-\frac{(x-x')^2}{2\sigma^2}}$$

is a PDK, since

$$K(x,x') = e^{\frac{xx'}{\sigma^2}} e^{-\frac{(x)^2}{2\sigma^2}} e^{-\frac{(x')^2}{2\sigma^2}}.$$

The first factor is the exponential kernel while the second and third factors are a tensor product.

Alternatively, we can conclude K is PDK through

$$e^{-\frac{t^2}{2\sigma^2}} = \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-i\omega t} e^{-\frac{\sigma^2\omega^2}{2}}\, d\omega.$$

Example: Laplace kernel. Let $\gamma > 0$ and $\mathcal{X} = \mathbb{R}$. Then, $K\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ defined as

$$K(x,x') = \frac{1}{2} e^{-\gamma|x-x'|}$$

is a PDK, since

$$\frac{1}{2}e^{-\gamma|t|} = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\omega t}\frac{\gamma}{\gamma^2+\omega^2}\,d\omega.$$

(Integral can be evaluated via contour integration.)

2.2 Reproducing kernel Hilbert space (RKHS)

Let $\mathcal{X}$ be a nonempty set (No further assumption on $\mathcal{X}$ yet). Let $\mathcal{H}$ be a Hilbert space of functions $f\colon \mathcal{X} \to \mathbb{R}$ equipped with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and induced norm $\|\cdot\|_{\mathcal{H}}$. (By definition, $\|f\|_{\mathcal{H}} = 0$ if and only if $f(x) = 0$ for all $x \in \mathcal{X}$. We say $K\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a reproducing kernel (RK) of $\mathcal{H}$ if

$$K(x,\cdot) \in \mathcal{H}, \qquad \forall\, x \in \mathcal{H},$$

and K has the *reproducing property*

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

If $\mathcal{H}$ has a RK, it is a *reproducing kernel Hilbert space* (RKHS).

Example: Band-limited functions. Let $B > 0$ and $\mathcal{X} = \mathbb{R}$. Let

$$\mathcal{H} = \left\{ f\colon \mathbb{R}\to\mathbb{R} \;\middle|\; \int_{\mathbb{R}\setminus[-B,B]} |\hat{f}(\omega)|^2\,d\omega = 0,\ \|f\|_{\mathcal{H}} < \infty,\ \hat{f} = \mathcal{F}[f],\ f = \mathcal{F}^{-1}[\hat{f}] \right\},$$

$$\langle f, g\rangle_{\mathcal{H}} = \int_{\mathbb{R}} f(x)g(x)\,dx = \frac{1}{2\pi}\int_{-B}^{B} \hat{f}(\omega)\overline{\hat{g}}(\omega)\,d\omega$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the forward and inverse Fourier transforms, be the Hilbert space of band-limited L^2 functions. Then, $\mathcal{H}$ is an RKHS with RK

$$K(x,x') = 2B\mathrm{sinc}(B(x-x')) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\omega x'}e^{i\omega x}\mathbf{1}_{[-B,B]}(\omega)\,d\omega.$$

To see why, note that

$$\widehat{K(x,\cdot)}(\omega) = e^{i\omega x}\mathbf{1}_{[-B,B]}(\omega),$$

so $K(x,\cdot) \in \mathcal{H}$ for all $x \in \mathbb{R}$, and

$$\langle f, K(x,\cdot)\rangle_{\mathcal{H}} = \frac{1}{2\pi}\int_{-B}^{B}\hat{f}(\omega)e^{-i\omega x}\,d\omega = \frac{1}{2\pi}\int_{\mathbb{R}}\hat{f}(\omega)e^{-i\omega x}\,d\omega = f(x),$$

so K has the reproducing property.

RKHSs can be equivalently defined by continuity of point evaluation.

Theorem 18. *Let $\mathcal{X}$ be a nonempty set. Let $\mathcal{H}$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$. $\mathcal{H}$ is an RKHS if and only if the evaluation functional L_x, defined as $L_x[f] = f(x)$, is bounded (continuous) for all $x \in \mathcal{X}$.*

Proof. Assume $\mathcal{H}$ is an RKHS. For any $x \in \mathcal{X}$,

$$\begin{aligned} |L_x[f]| &= |\langle f, K(x,\cdot)\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \|K(x,\cdot)\|_{\mathcal{H}}, \qquad \forall\, f \in \mathcal{H} \end{aligned}$$

and $\|K(x,\cdot)\|_{\mathcal{H}}$ is well defined and finite since $K(x,\cdot) \in \mathcal{H}$. So L_x is bounded.

Next, assume $L_x \colon \mathcal{H} \to \mathbb{R}$ is bounded in $\mathcal{H}$. By the Riesz representation theorem, there exists a $h_x \in \mathcal{H}$ such that

$$L_x[f] = \langle h_x, f\rangle_{\mathcal{H}}, \qquad \forall\, f \in \mathcal{H}.$$

Let $K(x,x') = h_x(x')$ for all $x, x' \in \mathcal{X}$. □

Interestingly, there is a one-to-one correspondence between PDKs and RKHSs. First, we establish uniqueness: if a $\mathcal{H}$ exists for a K, then it is unique; and if a K exists for a $\mathcal{H}$, then it is unique.

Theorem 19. *If $\mathcal{H}$ is an RKHS, its reproducing kernel $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is unique.*

Proof. Let K and K' be two RK of an RKHS $\mathcal{H}$. Then for any $x \in \mathcal{X}$,

$$\begin{aligned} \|K(x,\cdot) - K'(x,\cdot)\|_{\mathcal{H}}^2 &= \langle K(x,\cdot) - K'(x,\cdot), K(x,\cdot) - K'(x,\cdot)\rangle_{\mathcal{H}} \\ &= \langle K(x,\cdot), K(x,\cdot) - K'(x,\cdot)\rangle_{\mathcal{H}} - \langle K'(x,\cdot), K(x,\cdot) - K'(x,\cdot)\rangle_{\mathcal{H}} \\ &= K(x,x) - K'(x,x) - K(x,x) + K'(x,x) \\ &= 0. \end{aligned}$$

Therefore, $K = K'$. □

Theorem 20. *If $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel, its Hilbert space $\mathcal{H}$ is unique.*

Proof. Let $\mathcal{H}$ be an RKHS of a reproducing kernel K. Let

$$\mathcal{S} = \operatorname{span}\{K(x,\cdot) \,|\, x \in \mathcal{X}\}.$$

We claim $\overline{\mathcal{S}} = \mathcal{H}$, which holds if and only if 0 is the only element in $\mathcal{H}$ orthogonal to all vectors in $\mathcal{S}$. Indeed, if $h \in \mathcal{H}$ satisfies

$$\langle h, K(x,\cdot)\rangle = 0, \qquad \forall\, x \in \mathcal{X},$$

then $h(x) = 0$ for all $x \in \mathcal{X}$, by the reproducing property, and $h = 0$. Since, any RKHS of K is precisely characterized by $\overline{\mathcal{S}} = \mathcal{H}$, it is unique. □

In machine learning, we want to evaluate functions to make predictions, but point evaluations are not well-defined for L^p spaces, since elements of L^p spaces are equivalence classes of functions whose values may differ on a set of measure zero. Therefore, the requirements of RKHSs that the evaluation functional is continuous is, in some sense, a natural requirement.

We now complete the proof of the one-to-one correspondence between PDKs and RKHSs by showing existence: there exists a $\mathcal{H}$ exists for a K; and there exists a K for a $\mathcal{H}$.

Theorem 21 (Moore–Aronszajn Theorem)**.** *Let $\mathcal{X}$ be a nonempty set. Then $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PDK if and only if it is an RK of an RKHS $\mathcal{H}$.*

Proof. ($\Leftarrow$) Assume K is an RK of an RKHS $\mathcal{H}$. Then K is symmetric, since $K(x, x') = \langle K(x, \cdot), K(x', \cdot)\rangle_{\mathcal{H}} = \langle K(x', \cdot), K(x, \cdot)\rangle_{\mathcal{H}} = K(x', x)$ for all $x, x' \in \mathcal{X}$. Then for any $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$ and $c \in \mathbb{R}^N$, we have

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{N} c_i K(x_i, \cdot), \sum_{j=1}^{N} c_j K(x_j, \cdot) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{N} c_i K(x_i, \cdot) \right\|_{\mathcal{H}}^2 \\
&\geq 0.
\end{aligned}$$

So K is a PDK.

($\Rightarrow$) Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDK. Define $\mathcal{H}_0$ to be the (not necessarily complete) vector space

$$\begin{aligned}
\mathcal{H}_0 &= \operatorname{span}\{K(x, \cdot) \,|\, x \in \mathcal{X}\} \\
&= \left\{ \sum_{i=1}^{N} \alpha_i K(x_i, \cdot) \,\middle|\, N \in \mathbb{N},\, x_1, \ldots, x_N \in \mathcal{X},\, \alpha_1, \ldots, \alpha_N \in \mathbb{R} \right\}.
\end{aligned}$$

For

$$f = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot), \qquad g = \sum_{i=1}^{N'} \beta_i K(x'_i, \cdot),$$

define

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{N}\sum_{j=1}^{N'} \alpha_i \beta_j K(x_i, x'_j).$$

Clearly, $\langle \cdot, \cdot \rangle_{\mathcal{H}_0} \colon \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{R}$ is symmetric and bilinear. The value of $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is independent of the representation of f via $x_1, \ldots, x_N, \alpha_1, \ldots, \alpha_N$ and g via $x'_1, \ldots, x'_{N'}, \beta_1, \ldots, \beta_{N'}$, since

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{N} \alpha_i g(x_i) = \sum_{j=1}^{N'} \beta_j f(x'_j).$$

(So $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is well-defined.) Since K is a PDK, we have $\langle f, f \rangle_{\mathcal{H}_0} = \alpha^\intercal G \alpha \geq 0$, where $\alpha = (\alpha_1, \ldots, \alpha_N)$ and $G \in \mathbb{R}^{N \times N}$ is the kernel matrix for $x_1, \ldots, x_N$. So $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is a semi-inner product (it is an inner product, but we have so far shown that it is a semi-inner product.) so Cauchy–Schwartz inequality holds by Lemma 17. We do have the reproducing property

$$\langle f, K(x, \cdot) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{N} \alpha_i K(x_i, x) = f(x), \qquad \forall x \in \mathcal{X},\ f \in \mathcal{H}_0.$$

Therefore,

$$|f(x)| \leq |\langle f, K(x, \cdot) \rangle_{\mathcal{H}_0}| \leq \|f\|_{\mathcal{H}_0} \|K(x, \cdot)\|_{\mathcal{H}_0} \leq \|f\|_{\mathcal{H}_0} \sqrt{K(x, x)},$$

and $\|f\|_{\mathcal{H}_0} = 0$ implies $f(x) = 0$ for all $x \in \mathcal{X}$, i.e., $f = 0$. Therefore, $\mathcal{H}_0$ is a pre-Hilbert space (a vector space equipped with an inner product).

Finally, we complete the space to get $\mathcal{H}$ by considering Cauchy sequences in $\mathcal{H}_0$. We defer the arguments to Section 2.2.1. □

Lemma 17. *Cauchy–Schwartz inequality holds for semi-inner products.*

Proof. Let $\mathcal{V}$ be a real vector space and let $\langle \cdot, \cdot \rangle \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a semi-inner product, i.e., it is a bilinear map satisfying $\langle f, f \rangle \geq 0$. Then,

$$0 \leq \left\| \frac{\langle u, v \rangle}{\|u\|} u - \|u\| v \right\|^2 = \|u\|^2 \|v\|^2 - (\langle u, v \rangle)^2.$$

□

Example: Linear kernel. Let $\mathcal{X} = \mathbb{R}^d$ and

$$\mathcal{H} = \{ f_w(\cdot) = \langle w, \cdot \rangle_{\mathbb{R}^d},\ w \in \mathbb{R}^d \}, \qquad \langle f_w, f_v \rangle_{\mathcal{H}} = \langle w, v \rangle_{\mathbb{R}^d}$$

be the space of linear functions. The evaluation map

$$L_x[f_w] = f_w(x) = \langle w, x \rangle_{\mathbb{R}^d} = \langle f_w, f_x \rangle_{\mathcal{H}}$$

has the representation $f_x \in \mathcal{H}$, and the reproducing kernel is

$$K(x, x') = f_x(x') = \langle x, x' \rangle_{\mathbb{R}^d}.$$

Note that $f_\cdot \colon \mathbb{R}^d \to \mathcal{H}$ defines a feature map one can use to establish that K is PDK, since

$$K(x, x') = \langle f_x, f_{x'} \rangle_{\mathcal{H}}.$$

However, is it not the only feature map. Another example is $\phi \colon \mathbb{R}^d \to \mathbb{R}^d$ defined as $\phi(x) = x$, since

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^d} = \langle x, x' \rangle_{\mathbb{R}^d}.$$

While PDK and RKHS have one-to-one correspondences, feature maps $\phi \colon \mathcal{X} \to \mathcal{H}$ and kernels do not. However, we do have a one-to-one correspondence if we require $\phi \colon \mathcal{X} \to \mathcal{H}$ to map to an RKHS $\mathcal{H}$.

Example: Quadratic kernel. Let $\mathbf{S}^{d \times d}$ be the set of $d \times d$ symmetric matrices. For $S, R \in \mathbf{S}^{d \times d}$, define

$$\langle S, R \rangle_{\mathbf{S}^{d \times d}} = \sum_{i=1}^{d} \sum_{j=1}^{d} S_{ij} R_{ij} = \operatorname{Trace}(S^\intercal R) = \operatorname{Trace}(SR).$$

For $x \in \mathbb{R}^d$, we can use the "trace trick" to get

$$\begin{aligned} \langle S, xx^\intercal \rangle_{\mathbf{S}^{d \times d}} &= \operatorname{Trace}(Sxx^\intercal) \\ &= \operatorname{Trace}(x^\intercal S x) \\ &= x^\intercal S x. \end{aligned}$$

(The identity $\sum_i \sum_j S_{ij} R_{ij} = \operatorname{Trace}(S^\intercal R)$ holds even when S and R are not symmetric.)

Let $\mathcal{X} = \mathbb{R}^d$. For any $S \in \mathbf{S}^{d \times d}$, define $f_S \colon \mathbb{R}^d \to \mathbb{R}$ as

$$f_S(x) = x^\intercal S x = \langle S, xx^\intercal \rangle_{\mathbf{S}^{d \times d}}.$$

Let

$$\mathcal{H} = \{ f_S \mid S \in \mathbf{S}^{d \times d} \}, \qquad \langle f_S, f_{S'} \rangle_{\mathcal{H}} = \langle S, S' \rangle_{\mathbf{S}^{d \times d}}.$$

Then the evaluation map

$$L_x[f_S] = f_S(x) = \langle S, xx^\intercal \rangle_{\mathbf{S}^{d \times d}} = \langle f_S, f_{xx^\intercal} \rangle_{\mathcal{H}}$$

has the representation $f_{xx^\intercal} \in \mathcal{H}$, and the reproducing kernel is

$$K(x, x') = f_{xx^\intercal}(x') = \langle xx^\intercal, x'(x')^\intercal \rangle_{\mathbf{S}^{d\times d}} = (\langle x, x' \rangle_{\mathbb{R}^d})^2.$$

Note that while

$$\mathcal{H} = \operatorname{span}\{K(x, \cdot) \,|\, x \in \mathbb{R}^d\} = \operatorname{span}\{f_{xx^\intercal} \,|\, x \in \mathbb{R}^d\}$$

(the span is already complete without any need for completion), there exists $f \in \mathcal{H}$ such that $f \neq \alpha K(x, \cdot)$ for any $\alpha \in \mathbb{R}$ and $x \in \mathcal{X}$.

Example: Gaussian kernel. Let $\sigma > 0$ and $\mathcal{X} = \mathbb{R}$. Let

$$\mathcal{H} = \{f \colon \mathbb{R} \to \mathbb{R} \,|\, \|f\|_{\mathcal{H}} < \infty\}, \qquad \langle f, g \rangle_{\mathcal{H}} = \frac{1}{(2\pi)^{3/2}\sigma} \int_{\mathbb{R}} \hat{f}(\omega)\overline{\hat{g}}(\omega) e^{\frac{\sigma^2\omega^2}{2}}\, d\omega.$$

Then, $\mathcal{H}$ is an RKHS with the Gaussian RK

$$K(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}}.$$

To see why, note

$$\widehat{K(x, \cdot)}(\omega) = \sqrt{2\pi}\sigma e^{i\omega x} e^{-\frac{\sigma^2\omega^2}{2}},$$

so

$$\|K(x, \cdot)\|_{\mathcal{H}}^2 = \frac{\sigma}{(2\pi)^{1/2}} \int_{\mathbb{R}} e^{-\frac{\sigma^2\omega^2}{2}}\, d\omega < \infty$$

and

$$\langle f, K(x, \cdot) \rangle_{\mathcal{H}} = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega) e^{-i\omega x}\, d\omega = f(x).$$

One can also show that $\mathcal{H}$ is a class of all functions in L^2 (so Fourier and inverser Fourier transforms are well-defined) that are infinitely differentiable with all derivatives in L^2.

Example: Exponential kernel. Let $\mathcal{X} = \mathbb{R}$. Consider the exponential kernel

$$K(x, x') = e^{xx'}.$$

While we know that

$$\mathcal{H} = \overline{\operatorname{span}\{K(x, \cdot) = e^{x\cdot} \,|\, x \in \mathbb{R}\}},$$

a nice analytical characterization of $\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ seems to be unknown. Given a PDK K, it is not always straightforward to characterize the corresponding RKHS $\mathcal{H}$, and vice versa.

By now, it should be clear that $\phi\colon \mathbb{R} \to \mathcal{H}$ defined by $\phi(x) = K(x, \cdot)$ is a feature map we can use to establish that K is PDK, since

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

However, another feature map (mapping to a Hilbert space that is not an RKHS) is $\psi\colon \mathbb{R} \to \ell^2$

$$\psi(x) = (1, x/\sqrt{1!}, x^2/\sqrt{2!}, x^3/\sqrt{3!}, \dots) \in \ell^2,$$

since

$$K(x, x') = \langle \psi(x), \psi(x') \rangle_{\ell^2} = \sum_{i=0}^{\infty} \frac{(xx')^i}{i!} = e^{xx'}.$$

2.2.1 Completion argument of Moore–Aronszajn

We now complete the completion argument of the Moore–Aronszajn theorem.

Pointwise convergence and definition of $\mathcal{H}$. Let $\mathcal{H}_0$ be the pre-Hilbert space as constructed in the initial part of the proof of Theorem 21. Let $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ be a Cauchy sequence with respect to the $\|\cdot\|_{\mathcal{H}_0}$-norm. For any $x \in \mathcal{X}$,

$$\begin{aligned} |f_m(x) - f_n(x)| &= |\langle f_m - f_n, K(x, \cdot) \rangle_{\mathcal{H}_0}| \\ &\leq \|f_m - f_n\|_{\mathcal{H}_0} \|K(x, \cdot)\|_{\mathcal{H}_0} \\ &= \|f_m - f_n\|_{\mathcal{H}_0} \sqrt{K(x, x)} \\ &\to 0 \end{aligned}$$

as $\min\{m, n\} \to \infty$. So, for all $x \in \mathcal{X}$, $\{f_k(x)\}_{k\in\mathbb{N}} \subset \mathbb{R}$ is a Cauchy sequence and converges to a limit. We define $f_\infty\colon \mathcal{X} \to \mathbb{R}$ to be the pointwise limit of $\{f_k\}_{k\in\mathbb{N}}$, i.e.,

$$f_\infty(x) = \lim_{k\to\infty} f_k(x).$$

We define $\mathcal{H}$ as the space of all pointwise limits of Cauchy sequences in $\mathcal{H}_0$. Clearly, $\mathcal{H}$ is a vector space. Moreover, $\mathcal{H}_0 \subseteq \mathcal{H}$, since for any $f \in \mathcal{H}_0$, the Cauchy sequence $f_k = f$ for all k has the limit f.

Definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Let $f_\infty, g_\infty \in \mathcal{H}$ with Cauchy sequences $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ and $\{g_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ respectively converging to them. Define

$$\langle f_\infty, g_\infty \rangle_{\mathcal{H}} = \lim_{k\to\infty} \langle f_k, g_k \rangle_{\mathcal{H}_0}.$$

For this definition to be well defined, the limit must exist and the limit must not depend on the Cauchy sequence converging to $f_\infty, g_\infty \in \mathcal{H}$. First,

$$\begin{aligned}
|\langle f_m, g_m\rangle_{\mathcal{H}_0} - \langle f_n, g_n\rangle_{\mathcal{H}_0}| &= |\langle f_m - f_n, g_m\rangle_{\mathcal{H}_0} - \langle f_n, g_n - g_m\rangle_{\mathcal{H}_0}| \\
&\leq |\langle f_m - f_n, g_m\rangle_{\mathcal{H}_0}| + |\langle f_n, g_n - g_m\rangle_{\mathcal{H}_0}| \\
&\leq \|f_m - f_n\|_{\mathcal{H}_0}\|g_m\|_{\mathcal{H}_0} + \|f_n\|_{\mathcal{H}_0}\|g_n - g_m\|_{\mathcal{H}_0} \\
&\to 0
\end{aligned}$$

as $\min\{m, n\} \to \infty$. (Note that $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ and $\{g_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ are bounded since they are Cauchy.) Next, let $\{f'_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ and $\{g'_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ also be Cauchy sequences respectively converging to f_∞ and g_∞. Then

$$\begin{aligned}
|\langle f_n, g_n\rangle_{\mathcal{H}_0} - \langle f'_n, g'_n\rangle_{\mathcal{H}_0}| &= |\langle f_n - f'_n, g_n\rangle_{\mathcal{H}_0} - \langle f'_n, g'_n - g_n\rangle_{\mathcal{H}_0}| \\
&\leq |\langle f_n - f'_n, g_n\rangle_{\mathcal{H}_0}| + |\langle f'_n, g'_n - g_n\rangle_{\mathcal{H}_0}| \\
&\leq \|f_n - f'_n\|_{\mathcal{H}_0}\|g_n\|_{\mathcal{H}_0} + \|f'_n\|_{\mathcal{H}_0}\|g'_n - g_n\|_{\mathcal{H}_0} \\
&\to 0
\end{aligned}$$

as $n \to \infty$.

$\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product. That $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is symmetric and bilinear is clear. Also, $\|\cdot\|_{\mathcal{H}}$ is nonnegative, since

$$\|f_\infty\|_{\mathcal{H}} = \lim_{k\to\infty} \|f_k\|_{\mathcal{H}_0} \geq 0$$

for $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ converging to f_∞. For $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ to be an inner product on $\mathcal{H}$, it remains to verify positive definiteness of $\|\cdot\|_{\mathcal{H}}$, i.e., that $\|f_\infty\|_{\mathcal{H}} = 0$ only if and only if $f_\infty(x) = 0$ for all $x \in \mathcal{X}$.

If $f_\infty = 0$, then $\|f_\infty\|_{\mathcal{H}} = 0$ since $0 \in \mathcal{H}_0$ and $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ with $f_k = 0$ converges to $f_\infty = 0$. Conversely, assume $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ converges to f_∞ and $\|f_\infty\|_{\mathcal{H}} = 0$. Then, for any $x \in \mathcal{X}$,

$$|f_\infty(x)| = \left|\lim_{k\to\infty} f_k(x)\right| = \left|\lim_{k\to\infty} \langle f_k, K(x,\cdot)\rangle_{\mathcal{H}_0}\right| \leq \lim_{k\to\infty} \|f_k\|_{\mathcal{H}_0}\|K(x,\cdot)\|_{\mathcal{H}_0}.$$

Since $\|f_k\|_{\mathcal{H}_0} \to \|f\|_{\mathcal{H}} = 0$, we conclude $f_\infty(x) = 0$ for all $x \in \mathcal{X}$.

$\mathcal{H}$ is complete. While Cauchy sequences in $\mathcal{H}_0$ have limits $\mathcal{H}$ by definition, it remains to establish that Cauchy sequences in $\mathcal{H}$ have a limit in $\mathcal{H}$. We use the standard argument that the set of all equivalence classes of Cauchy sequences is complete.

Let $f_\infty^{(1)}, f_\infty^{(2)}, \ldots$ be a Cauchy sequence in $\mathcal{H}$, and let $\{f_k^{(1)}\}_{k\in\mathbb{N}}, \{f_k^{(2)}\}_{k\in\mathbb{N}}, \ldots$ be Cauchy sequences in $\mathcal{H}_0$ with respective limits $f_\infty^{(1)}, f_\infty^{(2)}, \ldots$. Let $\{k(j)\}_{j\in\mathbb{N}} \subseteq \mathbb{N}$ be a sequence such that $\|f_{k(j)}^{(j)} - f_\infty^{(j)}\| \to 0$ as $j \to \infty$. Then

$$\begin{aligned}\|f_{k(i)}^{(i)} - f_{k(j)}^{(j)}\|_{\mathcal{H}_0} &= \|f_{k(i)}^{(i)} - f_{k(j)}^{(j)}\|_{\mathcal{H}} \\ &\leq \|f_{k(i)}^{(i)} - f_\infty^{(i)}\|_{\mathcal{H}} + \|f_\infty^{(i)} - f_\infty^{(j)}\|_{\mathcal{H}} + \|f_\infty^{(j)} - f_{k(j)}^{(j)}\|_{\mathcal{H}} \\ &\to 0\end{aligned}$$

as $\min\{i, j\} \to \infty$. Therefore, $\{f_{k(j)}^{(j)}\}_{j\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{H}_0$ and it has a limit $\mathbf{f} \in \mathcal{H}$. Finally,

$$\|\mathbf{f} - f_\infty^{(j)}\|_{\mathcal{H}} \leq \|\mathbf{f} - f_{k(j)}^{(j)}\|_{\mathcal{H}} + \|f_{k(j)}^{(j)} - f_\infty^{(j)}\|_{\mathcal{H}} \to 0$$

as $j \to \infty$. Since the Cauchy sequence $f_\infty^{(1)}, f_\infty^{(2)}, \ldots$ in $\mathcal{H}$ converges to a limit $\mathbf{f}$ in $\mathcal{H}$, we conclude $\mathcal{H}$ is complete.

K is an RK for $\mathcal{H}$. We have established that K has the reproducing property for $\mathcal{H}_0$ and that $K(x, \cdot) \in \mathcal{H}_0 \subseteq \mathcal{H}$ for all $x \in \mathcal{X}$. It remains to show that K has the reproducing property for all of $\mathcal{H}$. Let $f_\infty \in \mathcal{H}$ and let $\{f_k\}_{k\in\mathbb{N}} \subset \mathcal{H}_0$ be a Cauchy sequence converging to f_∞. Then

$$\underbrace{f_k(x)}_{\to f_\infty(x)} = \underbrace{\langle f_k, K(x, \cdot)\rangle_{\mathcal{H}_0}}_{\to \langle f_\infty, K(x,\cdot)\rangle_{\mathcal{H}}}.$$

□

2.2.2 Discussion

RKHS norm quantifies smoothness. The norm of a function in an RKHS controls how fast the function varies over $\mathcal{X}$ with respect to the (pseudo-)metric d_K, defined as below. Alternatively, one says, $\|f\|_{\mathcal{H}}$ quantifies the "smoothness" of f. In the context of machine learning and optimization, "smoothness" often refers to the variation of the function, and does not directly refer to (infinite) differentiability. Specifically, for $f \in \mathcal{H}$,

$$\begin{aligned}|f(x) - f(x')| &= |\langle f, K(x, \cdot) - K(x', \cdot)\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}}\|K(x, \cdot) - K(x', \cdot)\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}} d_K(x, x'),\end{aligned}$$

so f is $\|f\|_{\mathcal{H}}$-Lipschitz continuous as a map from $(\mathcal{X}, d_K)$ to $(\mathbb{R}, |\cdot|)$.

2.3 Kernel trick in shallow learning

Many classical shallow learning techniques can be enhanced through the kernel trick. In particular, they can learn non-linear decision boundaries (despite being linear in the feature vectors) and can learn arbitrary decision boundaries if the kernel is "universal" (the estimator is statistically consistent). The full scope of the kernel trick is beyond the scope of this course. Rather, we will briefly present the core idea, the kernel trick.

The canonical example of this is logistic regression, where we have data $X \in \mathbb{R}^d$ and label $Y \in \{-1, +1\}$ and we solve

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}[\ell(Y\theta^\intercal X)],$$

with

$$\ell(z) = \log(1 + \exp(-z)).$$

The idea is that, once trained, $\text{sign}(\theta^\intercal X)$ will predict the corresponding label Y. In the following, we consider a slightly generalized formulation.

In the following, we consider the one-pass SGD setup, where each data point is used only once in training. The finite-sum formulation, which occurs more often in practice, will be discussed later.

Basic SGD. Let $\mathcal{X} \in \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\theta \in \mathbb{R}^d$, and $h_\theta(x) = \theta^\intercal x$. (So $h_\theta \colon \mathbb{R}^d \to \mathbb{R}$.) Let P be a probability distribution for data-label (X, Y) pairs. Consider the optimization problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}\left[\ell(h_\theta(X); Y)\right].$$

The idea is that the loss function ℓ is chosen appropriately so that the optimized $h_\theta(X)$ serves as a predictor for Y. (For logistic regression, $\text{sign}(h_\theta(X))$ predicts the corresponding label $Y \in \{-1, +1\}$.)

Stochastic gradient descent (SGD) uses IID samples of data-label pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N) \sim P$ to execute

$$\begin{aligned}
\theta^{k+1} &= \theta^k - \alpha_{k+1} \nabla_{\theta^k} \ell(h_\theta(X_{k+1}); Y_{k+1}) \\
&= \theta^k - \underbrace{\alpha_{k+1} \ell'(h_{\theta^k}(X_{k+1}); Y_{k+1})}_{=\beta_k} X_{k+1} \\
&= \theta^k - \beta_{k+1} X_{k+1}
\end{aligned}$$

for $k = 0, \ldots, N-1$, where $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ are learning rates and ℓ' denotes the 1-dimensional derivative with respect to the first input. For the sake

of simplicity, consider the starting point $\theta^0 = 0$. The gradient computation follows from the chain rule and

$$D_\theta(h_\theta(X_{k+1})) = D_\theta(\theta^\intercal X_{k+1}) = X_{k+1}.$$

This approach can be generalized/enhanced through the use of a feature map.

Feature vector SGD. Next, let $\mathcal{X}$ be a nonempty set, $\phi\colon \mathcal{X} \to \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\theta \in \mathbb{R}^d$, and $h_\theta(x) = \theta^\intercal \phi(x)$. (So $h_\theta\colon \mathcal{X} \to \mathbb{R}$.) Consider the optimization problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}\left[\ell(h_\theta(X); Y)\right].$$

SGD with the feature map uses IID samples of data-label pairs to execute

$$\begin{aligned}
\theta^{k+1} &= \theta^k - \alpha_{k+1} \nabla_\theta \ell(h_{\theta^k}(X_{k+1}); Y_{k+1}) \\
&= \theta^k - \underbrace{\alpha_{k+1} \ell'(h_{\theta^k}(X_{k+1}); Y_{k+1})}_{=\beta_{k+1}} \phi(X_{k+1}) \\
&= \theta^k - \beta_{k+1} \phi(X_{k+1}).
\end{aligned}$$

The only difference with the prior formulation is that that all instances of X_{k+1} have been replaced with $\phi(X_{k+1})$.

2.3.1 Feature maps

When performing machine learning with data X, it is often advantageous to compute features derived from the data X and provide it to the machine learning algorithm. (In the past, such features were often hand-engineered, while in modern times, these features are often learned via deep learning.)

The *feature map* $\phi\colon \mathcal{X} \to \mathbb{R}^d$ maps data to its features. A feature map with a large feature set can allow the machine learning system to learn more complicated decision boundaries, but the increased number of features will incur a computational cost. Ideally, we want the feature map to bee one that is sufficiently high-diemensional (even infinite-dimensional) while having a nice kernelized form. We will return to this point when we discuss the kernel trick.

Example: Linear kernel. Let $\mathcal{X} = \mathbb{R}$. The linear kernel

$$K(x, x') = xx'$$

arises from the feature map $\phi\colon \mathbb{R} \to \mathbb{R}$ defined as

$$\phi(x) = x.$$

This feature map essentially corresponds to using the learning algorithm directly on the data without creating any new features.

Example: Quadratic kernel. Let $\mathcal{X} = \mathbb{R}$. The quadratic kernel

$$K(x, x') = (xx' + 1)^2 = x^2(x')^2 + 2xx' + 1.$$

A feature map for this kernel is $\phi\colon \mathbb{R} \to \mathbb{R}^3$ defined as

$$\phi(x) = \begin{bmatrix} x^2 \\ \sqrt{2}x \\ 1 \end{bmatrix}.$$

This feature map provides x^2 (and the constant 1) as additional features for the learning algorithm. Using the feature x^2, the learning algorithm can learn non-linear decision boundaries.

Example: Exponential kernel. Let $\mathcal{X} = \mathbb{R}$. The exponential kernel

$$K(x, x') = e^{xx'}$$

arises from the feature map $\phi\colon \mathbb{R} \to \ell^2$

$$\phi(x) = (1, x/\sqrt{1!}, x^2/\sqrt{2!}, x^3/\sqrt{2!}, \dots) \in \ell^2.$$

Compared to the prior feature maps, this feature map contains all powers of x and therefore has the potential to learn a much more expressive function. However, now the feature vector is infinite-dimensional, so we cannot use the feature vector as is. The kernel trick is needed.

2.3.2 Kernel trick and kernel SGD

Hilbert space SGD. Let $\mathcal{X}$ be a nonempty set, $\mathcal{H}$ a Hilbert space (not necessarily an RKHS), $\phi\colon \mathcal{X} \to \mathcal{H}$, $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$, $\mathcal{Y} = \mathbb{R}$, $\theta \in \mathcal{H}$, and $h_\theta(x) = \langle \theta, \phi(x) \rangle_{\mathcal{H}}$. (So $h_\theta\colon \mathcal{X} \to \mathbb{R}$.) Consider the optimization problem

$$\underset{\theta \in \mathcal{H}}{\text{minimize}} \quad \mathbb{E}_{(X,Y)\sim P}\left[\ell(h_\theta(X); Y)\right].$$

SGD without the kernel trick would have to execute

$$\begin{aligned}
\theta^{k+1} &= \theta^k - \alpha_{k+1} \nabla_\theta \ell(h_{\theta^k}(X_{k+1}); Y_{k+1}) \\
&= \theta^k - \alpha_{k+1} \nabla_\theta \ell(\langle \theta^k, \phi(X_k) \rangle_{\mathcal{H}}; Y_k) \\
&= \theta^k - \underbrace{\alpha_{k+1} \ell'(h_{\theta^k}(X_{k+1}); Y_{k+1})}_{=\beta_{k+1}} \phi(X_{k+1}) \\
&= \theta^k - \beta_{k+1} \phi(X_{k+1}).
\end{aligned}$$

Again, let $\theta^0 = 0$. This is the same formulation as before, except that θ^k now resides in a Hilbert space. If $\mathcal{H}$ is infinite-dimensional, then this SGD, as is, is not implementable on a computer.

The issue is resolved by the *kernel trick*, which expresses an algorithm in terms of inner products of the feature map outputs without directly using the output of a feature map:

$$\begin{aligned} h_{\theta^k}(x) &= \langle \theta^k, \phi(x) \rangle_{\mathcal{H}}, \qquad \left(\text{since } \theta^k = -\sum_{i=1}^{k} \beta_i \phi(X_i) \right) \\ &= -\sum_{i=1}^{k} \beta_i \langle \phi(X_i), \phi(x) \rangle_{\mathcal{H}} \\ &= -\sum_{i=1}^{k} \beta_i K(X_i, x). \end{aligned}$$

Even if $\mathcal{H}$ is infinite-dimensional, if evaluation of K is implementable, then the Hilbert space SGD is implementable as follows:

$$\begin{aligned} h_{\theta^k}(X_{k+1}) &= -\sum_{i=1}^{k} \beta_i K(X_i, X_{k+1}). \\ \beta_{k+1} &= \alpha_{k+1} \ell'(h_{\theta^k}(X_{k+1}); Y_{k+1}) \\ \text{Storage} &\leftarrow (\beta_{k+1}, X_{k+1}) \end{aligned}$$

for $k = 0, \ldots, N-1$. Once training via SGD is complete, perform inference via

$$h_{\theta^N}(x) = -\sum_{k=1}^{N} \beta_k K(X_k, x).$$

RKHS SGD. Let $\mathcal{X}$ be a nonempty set, $\mathcal{H}$ an RKHS on $\mathcal{X}$ with RK K, and $\mathcal{Y} = \mathbb{R}$. Consider the optimization problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad \mathbb{E}_{(X,Y) \sim P} \left[\ell(f(X); Y) \right].$$

SGD in the RKHS is

$$\begin{aligned} f^{k+1} &= f^k - \alpha_k \nabla_f \ell(f^k(X_{k+1}); Y_{k+1}) \\ &= f^k - \alpha_k \nabla_f \ell(\langle f^k, K(X_{k+1}, \cdot) \rangle_{\mathcal{H}}; Y_{k+1}) \\ &= f^k - \underbrace{\alpha_k \ell'(f^k(X_{k+1}); Y_{k+1})}_{=\beta_k} K(X_{k+1}, \cdot) \\ &= f^k - \beta_k K(X_{k+1}, \cdot), \end{aligned}$$

where we set $f^0 = 0$.

The RKHS SGD can be implemented identically as before with

$$\begin{aligned} f^k(X_{k+1}) &= -\sum_{i=1}^{k} \beta_i K(X_i, X_{k+1}) \\ \beta_{k+1} &= \alpha_{k+1} \ell'(f^k(X_{k+1}); Y_{k+1}) \\ \text{Storage} &\leftarrow (\beta_{k+1}, X_{k+1}) \end{aligned}$$

and

$$f^N(x) = -\sum_{k=1}^{N} \beta_k K(X_k, x).$$

2.3.3 Finite-sum problems

Although we considered the one-pass setup for the sake of simplicity, it is actually more common to access a single data point multiple times throughout SGD (i.e., one often performs multiple epochs of training). In such a setup, it is more natural to think of minimizing a finite-sum objective, also called the empirical risk, rather than an expectation, also called the true risk. Discussing the statistical implications of minimizing the finite-sum, rather than the true expectation, is beyond the scope of this course. Here, we briefly show that the kernel trick applies in the same manner for the finite-sum setup as well.

Let, $\mathcal{X}$ be a nonempty set, $\mathcal{H}$ an RKHS on $\mathcal{X}$ with RK K, and $\mathcal{Y} = \mathbb{R}$. Let $X_1, \dots, X_N \in \mathcal{X}$ and $Y_1, \dots, Y_N \in \mathcal{Y}$ be fixed data and labels. Consider the optimization problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i); Y_i),$$

SGD on the finite-sum formulation samples random indices $i(1), i(2), \dots \sim \text{Uniform}\{1, \dots, N\}$ and performs the update

$$\begin{aligned} f^{k+1} &= f^k - \alpha_{k+1} \nabla_f \ell(f^k(X_{i(k+1)}); Y_{i(k+1)}) \\ &= f^k - \underbrace{\alpha_{k+1} \ell'(f^k(X_{i(k+1)}); Y_{i(k+1)})}_{=\beta_{k+1}} K(X_{i(k+1)}, \cdot) \\ &= f^k - \beta_{k+1} K(X_{i(k+1)}, \cdot) \end{aligned}$$

for $k = 0, 1, \ldots, K-1$. Again, let $f^0 = 0$. The RKHS SGD can be implemented identically as before with

$$\begin{aligned} f^k(X_{i(k+1)}) &= -\sum_{j=1}^{k} \beta_j K(X_{i(j)}, X_{i(k+1)}) \\ \beta_{k+1} &= \alpha_{k+1} \ell'(f^k(X_{i(k+1)}); Y_{i(k+1)}) \\ \text{Storage} &\leftarrow (\beta_{k+1}, i(k+1)) \end{aligned}$$

and

$$f^K(x) = \sum_{k=1}^{K} \beta_k K(X_{i(k)}, x).$$

2.3.4 Representer theorem

Interestingly, both RKHS SGDs in the one-pass and finite-sum formulations, produce solutions within

$$\operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N).$$

Is this optimal? An obvious answer is that the solution is not optimal. Since SGD converges to the optimal solution (under certain assumptions) but does not arrive at a solution in a finite number of iterations, the solutions produced by the RKHS SGDs are not optimal.

For the finite-sum setup, however, restricting the search to within $\operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N)$ is optimal, since an optimal solution provably lies within the said subspace.

Theorem 22 (Representer theorem)**.** *Let $\mathcal{X}$ be a nonempty set, $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a PDK, $\mathcal{H}$ the corresponding RKHS, $X_1, \ldots, X_N \in \mathcal{X}$, and $Y_1, \ldots, Y_N \in \mathbb{R}$. Consider the optimization problem*

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad L(\{(X_i, Y_i, f(X_i))\}_{i=1}^N) + Q(\|f\|_{\mathcal{H}})$$

where $Q \colon \mathbb{R}_+ \to \mathbb{R}$ is a strictly increasing function. Then, if a minimizer exists, any minimizer must be in

$$\operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N).$$

Proof. Let

$$\mathcal{S} = \operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N) \subseteq \mathcal{H}.$$

In homework 3, you are to show that $f \in \mathcal{S}^\perp$ implies $f(X_i) = 0$ for all $i = 1, \ldots, N$.

Let $f^\star$ be a minimizer. Let

$$f_\star = s + t$$

such that $s \in \mathcal{S}$ and $t \in \mathcal{S}^\perp$. Then

$$L(\{(X_i, Y_i, f^\star(X_i))\}_{i=1}^N) = L(\{(X_i, Y_i, s(X_i))\}_{i=1}^N)$$

while

$$Q(\|f^\star\|_{\mathcal{H}}) = Q\left(\sqrt{\|s\|_{\mathcal{H}}^2 + \|t\|_{\mathcal{H}}^2}\right) \geq Q(\|s\|_{\mathcal{H}}),$$

where equality holds if and only if $t = 0$. Since $f^\star$ is assumed to be a minimizer, we conclude $t = 0$. □

In the absence of a regularizer, we have a non-strict version of the representer theorem.

Theorem 23 (Non-strict representer theorem)**.** *Let $\mathcal{X}$ be a nonempty set, $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a PDK, $\mathcal{H}$ the corresponding RKHS, $X_1, \ldots, X_N \in \mathcal{X}$, and $Y_1, \ldots, Y_N \in \mathbb{R}$. Consider the optimization problem*

$$\underset{f \in \mathcal{H}}{\text{minimize}} \quad L(\{(X_i, Y_i, f(X_i))\}_{i=1}^N) + Q(\|f\|_{\mathcal{H}})$$

where $Q \colon \mathbb{R}_+ \to \mathbb{R}$ is a non-decreasing function. (So $Q = 0$ is possible.) Then, if a minimizer exists, there is a minimizer in

$$\operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N).$$

Proof. Homework exercise. □

The representer theorem tells us that one can find a global optimum in

$$\operatorname{span}(\{K(X_i, \cdot)\}_{i=1}^N)$$

for the finite sum setup. Therefore, there are kernel methods that parameterize the solution into the form

$$f = \sum_{k=1}^N \beta_k K(X_k, \cdot)$$

and then optimize over $\beta_1, \ldots, \beta_N$ using optimization methods such as SGD, or even things like Newton's method.

2.3.5 Kernel ridge regression

Before starting the main content, quickly establish the following identity.

Lemma 18 (Push-through identity)**.** *Let $\gamma > 0$, $U \in \mathbb{R}^{m\times n}$, and $V \in \mathbb{R}^{n\times m}$. Then*

$$(\gamma I + UV)^{-1}U = U(\gamma I + VU)^{-1},$$

assuming $(\gamma I + UV)$ is invertible.

Proof. Clearly,

$$U(\gamma I + VU) = (\gamma I + UV)U.$$

Left-multiply $(\gamma + UV)^{-1}$ and right-multiply $(\gamma + VU)^{-1}$. □

Let $\mathcal{X}$ be a nonempty set. Let $X_1, \dots, X_N \in \mathcal{X}$, $Y_1, \dots, Y_N \in \mathbb{R}$, $\phi\colon \mathcal{X} \to \mathbb{R}^d$, and $\lambda > 0$. Let

$$\Phi = \begin{bmatrix} \phi(X_1)^\intercal \\ \phi(X_2)^\intercal \\ \vdots \\ \phi(X_N)^\intercal \end{bmatrix} \in \mathbb{R}^{N\times d}, \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \in \mathbb{R}^N.$$

Consider the *ridge regression*[2] problem

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^{N}(h_\theta(X_i) - Y_i)^2 + \lambda\|\theta\|^2,$$

where $h_\theta\colon \mathcal{X} \to \mathbb{R}$ is defined as $h_\theta(x) = \phi(x)^\intercal\theta$. Equivalently,[3] we write

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{N}\|\Phi\theta - Y\|^2 + \lambda\|\theta\|^2.$$

Because the objective function is convex, the solution $\theta^\star$ is found by setting the gradient to 0

$$0 = \frac{2}{N}\Phi^\intercal(\Phi\theta^\star - Y) + 2\lambda\theta^\star,$$

[2]Regression with ℓ^2-regularization is referred to as ridge regression in classical statistics.

[3]Linear regression is an instance of the finite-sum formulation and its goal is to obtain a prediction function $h_{\theta^\star}$ (which is linear in θ but need not be linear in x) rather than to obtain the parameters θ.

which solves to

$$\begin{aligned}\theta^\star &= \underbrace{(\Phi^\intercal\Phi + \lambda N I)^{-1}}_{d\times d}\underbrace{\Phi^\intercal Y}_{d\times 1} \\ &= \underbrace{\Phi^\intercal}_{d\times N}\underbrace{(\Phi\Phi^\intercal + \lambda N I)^{-1}}_{N\times N}\underbrace{Y}_{N\times 1} \\ &= \Phi^\intercal \underbrace{(G + \lambda N I)^{-1} Y}_{=\varphi^\star\in\mathbb{R}^N},\end{aligned}$$

where we used the kernel matrix $G \in \mathbb{R}^{N\times N}$

$$G_{ij} = \phi(X_i)^\intercal \phi(X_j)$$

and the push-through identity. Once "training" is complete, i.e., $\theta^\star$ has been computed, we make predictions on new data $x \in \mathcal{X}$ with

$$\begin{aligned}h_{\theta^\star}(\cdot) &= \phi(\cdot)^\intercal \theta^\star \\ &= \sum_{i=1}^N \varphi_i^\star K(\cdot, X_i).\end{aligned}$$

Next, consider the same linear regression setup with the prediction function in an RKHS as the explicit optimization variable. Let $X_1, \dots, X_N \in \mathcal{X}$, $Y_1, \dots, Y_N \in \mathbb{R}$, $\lambda > 0$, $K\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a PDK, and $\mathcal{H}$ the corresponding RKHS. Consider the *kernel ridge regression* problem

$$\underset{f\in\mathcal{H}}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^N (f(X_i) - Y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

When $\mathcal{H}$ is infinite dimensional, this is an infinite-dimensional optimization problem. By the representer theorem, a minimizer has the expression

$$f(x) = \sum_{j=1}^N \varphi_j K(x, X_j),$$

so we plug this form in to get a finite-dimensional optimization problem

$$\underset{\varphi\in\mathbb{R}^N}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^N \left(\sum_{j=1}^N \varphi_j K(X_j, X_i) - Y_i\right)^2 + \lambda \left\|\sum_{j=1}^N \varphi_j K(X_j, \cdot)\right\|_{\mathcal{H}}^2.$$

Using the kernel matrix $G \in \mathbb{R}^{N \times N}$, we equivalently write

$$\underset{\varphi \in \mathbb{R}^N}{\text{minimize}} \quad \frac{1}{N}\|G\varphi - Y\|^2 + \lambda \varphi^\intercal G \varphi.$$

The solution is found by setting the gradient to 0

$$0 = \frac{2}{N} G(G\varphi^\star - Y) + 2\lambda G \varphi^\star$$

and solves to

$$\varphi^\star = (G + \lambda N I)^{-1} Y.$$

(For the sake of simplicity, let us assume G is invertible. When G is not invertible, $\varphi^\star$ is a solution, but not the unique one. More on this in the homework assignment.) So, we have

$$f^\star(\cdot) = \sum_{j=1}^{N} \varphi_j^\star K(\cdot, X_j).$$

This is exactly the same prediction function as before, except that we did not need to have a finite-dimensional feature map.

Kernelized implementation. To conclude, given $X_1, \ldots, X_N \in \mathcal{X}$, $Y_1, \ldots, Y_N \in \mathbb{R}$, $\lambda > 0$, and a PDK $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can implement kernel ridge regression in a kernelized manner by forming the kernel matrix $G \in \mathbb{R}^{N \times N}$ (requires $N(N+1)/2$ evaluations of $K(\cdot, \cdot)$ but no need to explicitly form a feature vector) and perform linear algebra computations to solve

$$\varphi^\star = (G + \lambda N I)^{-1} Y.$$

Then, prediction on new data $x \in \mathcal{X}$ can be made with

$$f^\star(x) = \sum_{j=1}^{N} K(x, X_j) \varphi_j.$$

When $\lambda = 0$. When $\lambda = 0$, i.e., when there is no ℓ^2-regularizer, the same line of analysis and derivation can be carried out with the Moore–Penrose pseudoinverse. In particular, one arrives at

$$\varphi^\star = G^\dagger Y.$$

2.3.6 RKHS with finite-dimensional feature vector and and corresponding 2-layer neural networks

Let $\phi\colon \mathcal{X} \to \mathbb{R}^d$ and write

$$\phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_d(x) \end{bmatrix}.$$

Assume $\phi_1, \ldots, \phi_d$ are linearly independent as functions. Consider $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K(x, x) = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^d}.$$

Let

$$\mathcal{H} = \operatorname{span}\{\phi_k\}_{k=1}^d.$$

For

$$f = \sum_{k=1}^d \alpha_k \phi_k, \qquad g = \sum_{k=1}^d \beta_k \phi_k,$$

define the inner product

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{k=1}^d \alpha_k \beta_k.$$

(If $\phi_1, \ldots, \phi_d$ are not linearly independent, the inner product is not uniquely defined.) It is relatively straightforward to show that $\mathcal{H}$ is a Hilbert space. We claim that $\mathcal{H}$ is the RKHS corresponding to K. We provide two separate justifications.

First, we provide a direct verification. Clearly,

$$K(x, \cdot) = \sum_{k=1}^d \underbrace{\phi_k(x)}_{=\gamma_k} \phi_k(\cdot) \in \operatorname{span}\{\phi_k\}_{k=1}^d = \mathcal{H}$$

for all $x \in \mathcal{X}$. Next,

$$\langle f, K(x, \cdot) \rangle_{\mathcal{H}} = \sum_{k=1}^d \alpha_k \gamma_k = \sum_{k=1}^d \alpha_k \phi_k(x) = f(x).$$

So we have the reproducing property. Therefore, K is the RK of $\mathcal{H}$.

On the other hand, the construction of the Moore–Aronszajn Theorem and Lemma 13 tells us

$$\begin{aligned}\mathcal{H} &= \operatorname{span}\{K(x,\cdot) \,|\, x \in \mathcal{X}\} \\ &= \operatorname{span}\left\{ K(x,\cdot) = \sum_{k=1}^{d} \phi_k(x)\phi_k(\cdot) \,|\, x \in \mathcal{X} \right\} \\ &= \operatorname{span}\{\phi_1, \ldots, \phi_d\}.\end{aligned}$$

The Moore–Aronszajn construction further tells us that for

$$f = \sum_{i=1}^{N} \tilde{\alpha}_i K(x_i,\cdot), \qquad g = \sum_{j=1}^{N'} \tilde{\beta}_j K(x'_j,\cdot)$$

we have

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{N}\sum_{j=1}^{N'} \tilde{\alpha}_i\tilde{\beta}_j K(x_i, x'_j).$$

We can simplify as

$$f = \sum_{i=1}^{N} \tilde{\alpha}_i \sum_{k=1}^{d} \phi_k(x_i)\phi_k(\cdot) = \sum_{k=1}^{d} \underbrace{\sum_{i=1}^{N} \tilde{\alpha}_i \phi_k(x_i)}_{=\alpha_k} \phi_k(\cdot)$$

and

$$g = \sum_{j=1}^{N'} \tilde{\beta}_j \sum_{k=1}^{d} \phi_k(x'_j)\phi_k(\cdot) = \sum_{k=1}^{d} \underbrace{\sum_{j=1}^{N'} \tilde{\beta}_j \phi_k(x'_j)}_{=\beta_k} \phi_k(\cdot),$$

so

$$\begin{aligned}\langle f, g\rangle_{\mathcal{H}} &= \sum_{i=1}^{N}\sum_{j=1}^{N'} \tilde{\alpha}_i\tilde{\beta}_j \sum_{k=1}^{d} \phi_k(x_i)\phi_k(x'_j) \\ &= \sum_{k=1}^{d}\sum_{i=1}^{N} \tilde{\alpha}_i\phi_k(x_i) \sum_{j=1}^{N'} \tilde{\beta}_j\phi_k(x'_j) \\ &= \sum_{k=1}^{d} \alpha_k\beta_k.\end{aligned}$$

Absence of independence assumption. If we do not assume that $\phi_1, \dots, \phi_d$ are linearly independent as functions, then the RKHS corresponding to

$$K(x, x) = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^d}$$

becomes $\mathcal{H} = \operatorname{span}\{\phi_k\}_{k=1}^d$ with the so-called *variation norm*

$$\|f\|_{\mathcal{H}}^2 = \left(\begin{array}{ll} \underset{\theta \in \mathbb{R}^d}{\text{minimize}} & \sum_{k=1}^d \gamma_k^2 \\ \text{subject to} & f = \sum_{k=1}^d \gamma_k \phi_k \end{array} \right).$$

Connection to 2-layer neural networks. Let $\mathcal{X} = \mathbb{R}^d$. Let $\phi_1, \dots, \phi_N$ be defined as

$$\phi_k(x) = \sigma(a_k^\intercal x + b_k).$$

Then

$$\mathcal{H} = \left\{ \sum_{k=1}^N u_k \sigma(a_k^\intercal x + b_k) \,|\, u_1, \dots, u_N \in \mathbb{R} \right\},$$

i.e., $\mathcal{H}$ is the set of 2-layer nerual networks with hidden layer weights and biases fixed to $a_1, \dots, a_N$ and $b_1, \dots, b_N$. Performing kernel SGD or kernel ridge regression corresponds to training the output layer weights of a 2-layer neural network with the hidden layer weights and biases fixed (and not trained).

2.4 Kernel as linear operators

Let $\nu \in \mathcal{M}_+(\mathcal{X})$ and consider $L^2(\nu)$. (So ν is nonnegative.) For $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, define the linear functional

$$L_K[f] = \int_{\mathcal{X}} K(\cdot, x') f(x') \, d\nu(x')$$

for $f \colon \mathcal{X} \to \mathbb{R}$. Clearly, L_K is linear, but we need additional assumptions to ensure that the integral is well-defined. In particular, we will show that the following RHS

$$\int_{\mathcal{X}} |K(x, x') f(x')| \, d\nu(x') \le \|K(\cdot, x)\|_{L^2(\nu)} \|f(\cdot)\|_{L^2(\nu)}$$

is finite under appropriate assumptions.

2.4.1 Mercer kernel and Mercer's theorem

We say a PDK $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *Mercer kernel* if $\mathcal{X}$ is a nonempty a compact metric space (or more generally a compact Hausdorff space) and K is continuous as a function (with respect to the Borel topology).

If K is a Mercer kernel, then

$$L_K\colon L^2(\nu) \to L^2(\nu).$$

to see why, first note that

$$\begin{aligned}\|K(\cdot, x)\|_{L^2(\nu)}^2 &= \int_{\mathcal{X}} (K(x, x'))^2 \, d\nu(x') \\ &\leq \sup_{x,x' \in \mathcal{X}} (K(x, x'))^2 \nu(\mathcal{X}) < \infty.\end{aligned}$$

Therefore, for $f \in L^2(\nu)$, $L_K[f](x)$ is well defined (pointwise) for all $x \in \mathcal{X}$. It remains to show that $L_K[f] \in L^2(\nu)$. This follows from the fact that $L_K[f]$ is a bounded function:

$$\begin{aligned}|L_K[f](x)| &= \left| \int_{\mathcal{X}} K(x, x') f(x') \, d\nu(x') \right| \\ &\leq \|K(\cdot, x)\|_{L^2(\nu)} \|f(\cdot)\|_{L^2(\nu)} \\ &\leq \sqrt{\nu(\mathcal{X})} \sup_{x,x'} |K(x, x')| \|f(\cdot)\|_{L^2(\nu)} \\ &< \infty.\end{aligned}$$

Theorem 24. *Let $\mathcal{X}$ be a compact metric space (or more generally a compact Hausdorff space) and $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous positive definite Mercer kernel. Let $\nu \in \mathcal{M}_+(\mathcal{X})$ and assume $\operatorname{supp}(\nu) = \mathcal{X}$, i.e., $\nu(U) > 0$ for any nonempty open $U \subseteq \mathcal{X}$. Then there exists an eigenfunction expansion*

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x'),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\{\psi_i\}_{i=1}^{\infty}$ are continuous functions that are orthonormal in the $L^2(\nu)$ inner product. The convergence is absolute for each $x, x' \in \mathcal{X}$ and uniform on $\mathcal{X} \times \mathcal{X}$.

Proof outline. If K is a Mercer kernel, then $L_K\colon L^2(\nu) \to L^2(\nu)$ is a compact, self-adjoint, nonnegative bounded linear operator. We then appeal to the spectral theorem. □

Feature map via eigenfunctions. Assume a PDK $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has a series expansion

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$$

or, more concisely,

$$K = \sum_{i=1}^{\infty} \lambda_i \psi_i \otimes \psi_i$$

with nonnegative $\lambda_1, \lambda_2, \ldots$. Define $\phi\colon \mathcal{X} \to \ell^2$ with

$$\phi(x) = \begin{bmatrix} \sqrt{\lambda_1}\psi_1(x) \\ \sqrt{\lambda_2}\psi_2(x) \\ \vdots \end{bmatrix}.$$

Then ϕ is a feature map for K:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\ell^2}$$

(Since $K(x, x) < \infty$, we have $\phi(x) \in \ell^2$.)

RKHS. Assume a PDK $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has a series expansion

$$K = \sum_{i=1}^{\infty} \lambda_i \psi_i \otimes \psi_i$$

with nonnegative $\lambda_1, \lambda_2, \ldots$. Assume that $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and that $\{\psi_i\}_{i \in \mathbb{N}}$ is linearly independent as functions. Then the corresponding RKHS is

$$\mathcal{H} = \left\{ f = \sum_{i=1}^{\infty} \alpha_i \psi_i \,\middle|\, \|f\|_{\mathcal{H}} < \infty \right\},$$

and for

$$f = \sum_{i=1}^{\infty} \alpha_i \psi_i, \qquad g = \sum_{i=1}^{\infty} \beta_i \psi_i$$

the inner product is

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{\alpha_i \beta_i}{\lambda_i}.$$

Covariance kernel. Assume a PDK $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has the integral form

$$K(x, x') = \mathbb{E}_{w\sim P}[\psi(x; w)\psi(x'; w)] = \int_{\mathcal{W}} \psi(x; w)\psi(x'; w)dP(w),$$

where $w \in \mathcal{W}$ is a random variable. More concisely, we can view $\psi(\cdot; w)$ as random function (randomness determined by $w \in \mathcal{W}$) and write

$$K = \mathbb{E}_{\psi}[\psi \otimes \psi].$$

Assume a $\{\psi(\cdot; w)\}_{w\in\mathcal{W}}$ are linearly independent as functions. This implies a function

$$f = \int_{\mathcal{W}} \alpha(w)\psi(\cdot; w)\ dP(w)$$

is uniquely identified by α, in the sense that

$$\int_{\mathcal{W}} \alpha(w)\psi(\cdot; w)\ dP(w) = \int_{\mathcal{W}} \alpha'(w)\psi(\cdot; w)\ dP(w)$$

if and only if $\alpha = \alpha'$ P-almost everywhere.

Then the corresponding RKHS is

$$\mathcal{H} = \left\{ f = \int_{\mathcal{W}} \alpha(w)\psi(\cdot; w)\ dP(w) \,\middle|\, \|f\|_{\mathcal{H}} < \infty \right\},$$

and for

$$f = \int_{\mathcal{W}} \alpha(w)\psi(\cdot; w)\ dP(w), \qquad g = \int_{\mathcal{W}} \beta(w)\psi(\cdot; w)\ dP(w)$$

the inner product is

$$\langle f, g\rangle_{\mathcal{H}} = \int_{\mathcal{W}} \alpha(w)\beta(w)\ dP(\theta).$$

An example of interest to us is

$$K(x, x') = \mathbb{E}_{(a,b)\sim P}[\sigma(a^\intercal x + b)\sigma(a^\intercal x + b)].$$

In this case, the RKHS contains functions of the form

$$f(x) = \int \alpha(a, b)\sigma(a^\intercal x + b)\ dP(a, b),$$

which can be viewed as infinite-width 2-layer neural networks.

2.5 Matrix-valued PDKs and vector-valued RKHSs

Let $\mathcal{X}$ be a nonempty set. Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$. We say K is symmetric if $K(x,x') = (K(x',x))^\intercal$ for all $x, x' \in \mathcal{X}$. Then K is a *matrix-valued positive definite kernel* (mvPDK) if it is symmetric and if for all $N \in \mathbb{N}$, $x_1, \dots, x_N \in \mathcal{X}$, and $c_1, \dots, c_N \in \mathbb{R}^d$, we have

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i^\intercal K(x_i, x_j) c_j \geq 0.$$

Clearly, if $K_1\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ and $K_2\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ are mvPDKs, then $K_1 + K_2$ is as well. Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ and $k\colon (\mathcal{X} \times \{1,\dots,d\}) \times (\mathcal{X} \times \{1,\dots,d\}) \to \mathbb{R}$ such that

$$(K(x,x'))_{ij} = k((x,i),(x,j)).$$

Then K is a mvPDK if and only if k is a (scalar-valued) PDK. (This equivalence will be established in a homework assignment.)

Let $\mathcal{H}$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}^d$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and induced norm $\|\cdot\|_{\mathcal{H}}$. We say $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ is reproducing kernel (RK) for $\mathcal{H}$ if

$$K(\cdot,x)w \in \mathcal{H}, \qquad \forall\, x \in \mathcal{X},\, w \in \mathbb{R}^d.$$

and K has the reproducing property

$$\langle f, K(\cdot,x)w\rangle_{\mathcal{H}} = w^\intercal f(x), \qquad \forall\, x \in \mathcal{X},\, w \in \mathbb{R}^d.$$

If $\mathcal{H}$ has an RK, it is a vector-valued RKHS (vvRKHS).

Theorem 25. *There is a one-to-one correspondence between mvPDKs and vvRKHSs.*

Example: Kronecker product. Let k be a scalar PDK and $M \in \mathbb{R}^{d\times d}$ a symmetric positive semidefinite matrix. Then

$$K = k \otimes M,$$

defined as

$$K(x,x') = k(x,x')M$$

is a mvPDK. To see why, for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and $c_1, \ldots, c_N \in \mathbb{R}^d$, we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} c_i^\intercal K(x_i, x_j) c_j &= \sum_{i=1}^{N}\sum_{j=1}^{N} c_i^\intercal M c_j k(x_i, x_j) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} N_{ij} k(x_i, x_j) \\ &= \operatorname{Tr}(NG) \geq 0,\end{aligned}$$

where we use the fact that $N \in \mathbb{R}^{N \times N}$ defined as $N_{ij} = c_i^\intercal M c_j k$ is symmetric positive semidefinite and the fact that the inner product between two symmetric positive semidefinite is nonnegative. (In convex analysis, one says the PSD cone is self-dual.) Alternatively, and essentially equivalently, one can show that the Kronecker product between $M \in \mathbb{R}^{d \times d}$ and $G \in \mathbb{R}^{N \times N}$ is symmetric positive semidefinite.

Example: Outer product. Let $f \colon \mathcal{X} \to \mathbb{R}^d$. Then

$$K = f \otimes f$$

defined as

$$K(x, x') = f(x) f(x')^\intercal$$

is a mvPDK. To see why, for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and $c_1, \ldots, c_N \in \mathbb{R}^d$, we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} c_i^\intercal K(x_i, x_j) c_j &= \sum_{i=1}^{N}\sum_{j=1}^{N} (c_i^\intercal f(x_i))(f(x_j)^\intercal c_j) \\ &= \sum_{i=1}^{N} (c_i^\intercal f(x_i)) \sum_{j=1}^{N} (c_j^\intercal f(x_j)) \\ &\geq 0.\end{aligned}$$

2.5.1 Tensor products

The tensor product is an operation generally defined between two vectors and vector spaces. For the sake of notational simplicity, we will use special instances of this notation. We simply point out, but do not justify, that our notation is consistent with the general tensor product operation.

Let $f\colon \mathcal{X} \to \mathbb{R}^m$ and $g\colon \mathcal{Y} \to \mathbb{R}^n$ and

$$F = f \otimes g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{m \times n}$$

is defined as

$$F(x, y) = f(x)g(y)^\intercal.$$

In particular,

$$F = f \otimes f\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$$

is defined as

$$F(x, x') = f(x)f(x')^\intercal.$$

If $f\colon \mathcal{X} \to \mathbb{R}$ and $m \in \mathbb{R}^d$, then

$$F = f \otimes m\colon \mathcal{X} \to \mathbb{R}^d$$

is defined as

$$F(x) = f(x)m.$$

(The Kronecker kernel was defined with this notation.) Finally, if $\mu \in \mathcal{M}(\mathcal{X})$ and $\nu \in \mathcal{M}(\mathcal{Y})$, then

$$\mu \otimes \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$$

is defined as

$$(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$$

for measurable $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$

$$\int_{\mathcal{X} \times \mathcal{Y}} f(x, y) d(\mu \otimes \nu) = \int_{\mathcal{Y}} \int_{\mathcal{X}} f(x, y) d\mu(x) d\nu(y)$$

for (measurable and integrable) $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In particular, if μ and ν are probability measures respectively for random variables X and Y, then $\mu \otimes \nu$ is the probability measure for (X, Y) such that X and Y are independent random variables with respective marginal probability distributions μ and ν.

2.6 Random feature learning

2.6.1 Kernel approximation

Let $\mathcal{X} = \mathbb{R}^d$. Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined as

$$\begin{aligned} K(x, x') &= \int_{\mathcal{W}} e^{-iw^\intercal(x - x')} \, dP(w) \\ &= \int_{\mathcal{W}} (\cos(w^\intercal x) \cos(w^\intercal x') + \sin(w^\intercal x) \sin(w^\intercal x')) \, dP(w), \end{aligned}$$

for some real nonnegative probability measure P. Consider using M IID random features with $w_1, \dots, w_M \sim P$ and $b_1, \dots, b_M \sim \text{Uniform}[0, 2\pi]$:

$$\phi(x) = \frac{\sqrt{2}}{\sqrt{M}} \begin{bmatrix} \cos(w_1^\intercal x + b_1) \\ \cos(w_2^\intercal x + b_2) \\ \vdots \\ \cos(w_M^\intercal x + b_M) \end{bmatrix}.$$

Then

$$\begin{aligned} \hat{K}(x, x') \approx \langle \phi(x), \phi(x) \rangle_{\mathbb{R}^M} &= \frac{1}{M} \sum_{i=1}^{M} 2 \cos(w_i^\intercal x + b_i) \cos(w_i^\intercal x' + b_i) \\ &\overset{(*)}{\to} K(x, x'). \end{aligned}$$

Loosely speaking, $(*)$ follows from a law-of-large-number type of argument, since

$$\begin{aligned} \mathbb{E}_{w,b}[2 \cos(w^\intercal x + b) \cos(w^\intercal x' + b)] &= \mathbb{E}_{w,b}[\cos(w^\intercal x + b) \cos(w^\intercal x' + b) + \sin(w^\intercal x + b) \sin(w^\intercal x' + b)] \\ &= \mathbb{E}_w[\Re e^{-iw^\intercal (x - x')}] \\ &= \mathbb{E}_w[\cos(w^\intercal x) \cos(w^\intercal x') + \sin(w^\intercal x) \sin(w^\intercal x')] \\ &= K(x, x'), \end{aligned}$$

where the first equality follows from uniformity of b.

Precisely speaking, $(*)$ is in the following sense.

Theorem 26 (Informal)**.** *Let $\Omega \subset \mathbb{R}^d$ be compact. Let $\delta > 0$. With probability $1 - \delta$,*

$$\sup_{x, x' \in \Omega} \left| \hat{K}(x, x') - K(x, x') \right| \leq \textit{small}.$$

In practice, one would use $\hat{K}$ in place of K for computational efficiency. Kernel ridge regression or any kernelized method requires storing all $X_1, \dots, X_N$ and prediction with a new x requires N additional kernel evaluations. On the other hand, using $\hat{K}$ does not require storing all data and prediction only requires $\mathcal{O}(Md)$ operations, as we discuss now.

Kernelized learning with $\hat{K}$ is equivalent to estimating $\theta \in \mathbb{R}^M$ of the prediction function

$$f_\theta(x) = \sum_{i=1}^{M} \theta_i \cos(w_i^\intercal x + b),$$

since f_θ parameterizes all functions within the RKHS corresponding to $\hat{K}$. (The factor $\sqrt{2/M}$ is absorbed into θ). This is a 2-layer neural network with cosine activation functions. The hidden layer's weights and biases are randomly initialized and not trained. The "output layer weights" $\theta \in \mathbb{R}^M$ are trained.

More generally, consider a PDK $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with an integral form

$$K(x, x') = \mathbb{E}_{w \sim P}[\psi(x; w)\psi(x'; w)] = \int_{\mathcal{W}} \psi(x; w)\psi(x'; w)\, dP(w),$$

where $w \in \mathcal{W}$ is a random variable and P is a probability measure. Consider using M IID random features with $w_1, \ldots, w_M \sim P$:

$$\phi(x) = \frac{1}{\sqrt{M}} \begin{bmatrix} \psi(x; w_1) \\ \psi(x; w_2) \\ \vdots \\ \psi(x; w_M) \end{bmatrix}.$$

Then we can expect

$$\hat{K}(x, x') = \langle \phi(x), \phi(x'), \rangle_{\mathbb{R}^M} \to K(x, x')$$

by the law of large numbers. Kernelized learning with $\hat{K}$ is equivalent to estimating $\theta \in \mathbb{R}^M$ of the prediction function

$$f_\theta(x) = \sum_{i=1}^{M} \theta_i \psi(x; w_i).$$

This includes 2-layer neural network with activation function σ

$$f_\theta(x) = \sum_{i=1}^{M} \theta_i \sigma(a_i^\intercal x + b_i),$$

where $a_1, \ldots, a_M$ and $b_1, \ldots, b_M$ are initialized randomly and fixed.

2.6.2 Function approximation

However, approximating the kernel $\hat{K} \approx K$, in whatever sense, is not the end result we desire. The desired end result is to find a prediction function f_θ that either approximates the true unknown funciton $f_\star$ well, or attained "low risk".

Let $X_1, \ldots, X_N \in \mathcal{X}$ be IID samples from a probability measure $\mu \in \mathcal{M}_+(\mathcal{X})$. Let

$$R_{\text{emp}}[f] = \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - f^\star(X_i))^2$$

and

$$R_{\text{true}}[f] = \mathbb{E}_{X\sim\mu}[(f(X) - f^\star(X))^2] = \|f - f^\star\|^2_{L^2(\mu)}.$$

In machine learning, the goal is usually to find a prediction function f with small *true risk* $R_{\text{true}}[f]$. With finite data, this is accomplished by instead minimizing the *empirical risk* $R_{\text{emp}}[f]$.

Consider prediction functions of the following form. Let $\psi(x; w)$ be a *random feature function* such that $|\psi(x; w)| \leq 1$ for all $x \in \mathbb{R}^d$ and $w \in \mathcal{W}$. Let $P \in \mathcal{M}_+(\mathcal{W})$ be a probability measure and sample $w_1, \ldots, w_M \overset{\text{IID}}{\sim} P$ to form

$$f_\theta(x) = \sum_{i=1}^{M} \theta_i \psi(x; w_i).$$

Finally, we solve

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}} \quad R_{\text{emp}}[f_\theta] = \frac{1}{N}\sum_{i=1}^{N}(f_\theta(X_i) - f^\star(X_i))^2$$

to obtain the solution $\hat{\theta}$.

Since $\hat{\theta}$ minimizes the surrogate objective, rather than the true objective, there is no reason to expect $\hat{\theta}$ to minimize $R_{\text{true}}[f_\theta]$. However, we will establish that $\hat{\theta}$ is close to optimal (for the true objective) in the sense that

$$R_{\text{true}}[f_{\hat{\theta}}] \approx \inf_\theta R_{\text{true}}[f_\theta].$$

Theorem 27. *Let*

$$f_\star = \int_{\mathcal{W}} \psi(\cdot; w)\, dQ(w),$$

where $Q \in \mathcal{M}(\mathcal{W})$ is a (signed) measure absolutely continuous with respect to P.[4] *Let $\mu \in \mathcal{M}_+(\mathcal{X})$ be any probability measure. Let $\delta > 0$. Then, with probability $1 - \delta$, there exists a $\tilde{\theta} \in \mathbb{R}^M$ such that*

$$R_{\text{true}}[f_{\tilde{\theta}}] = \|f_{\tilde{\theta}} - f^\star\|^2_{L^2(\mu)} \leq \frac{\sup_w |dQ/dP(w)|^2}{M}\left(1 + \sqrt{2\log(1/\delta)}\right)^2.$$

Proof outline. Let

$$\hat{f}^\star(x) = \sum_{i=1}^{M} \underbrace{\frac{1}{M}\frac{dQ}{dP}(w_i)}_{=\tilde{\theta}_i} \psi(x; w_i),$$

[4]We sample from P while $f_\star$ is represented with Q. In practice, we do not know Q, so we cannot sample from it. The probabilistic argument by Jones requires sampling from Q and therefore is not algorithmically implementable.

where $\frac{dQ}{dP}$ is the Radon–Nikodym derivative. Then

$$\mathbb{E}[\hat{f}^\star] = f^\star, \qquad \mathbb{E}[\|\hat{f}^\star - f^\star\|^2] \leq \sup_w \left|\frac{dQ}{dP}(w)\right|^2.$$

The claim follows from considering the variance obtaining the probability $1-\delta$ guarantee with McDiarmid's inequality. □

Theorem 28 (Informal)**.** *Assume some additional conditions. Then, with probability $\delta > 0$, we have*

$$|R_{\text{true}}[f] - R_{\text{emp}}[f]| \leq \mathcal{O}(1/\sqrt{N}).$$

Proof outline. Proof uses the notion of Rademacher complexity. □

We now conclude

$$\begin{aligned} R_{\text{true}}[f_{\hat{\theta}}] &\approx R_{\text{emp}}[f_{\hat{\theta}}] \\ &\leq R_{\text{emp}}[f_{\tilde{\theta}}] \\ &\approx R_{\text{true}}[f_{\tilde{\theta}}] \\ &= \text{small}, \end{aligned}$$

where the chain of reasoning follows from Theorem 28, by definition of $\hat{\theta}$, Theorem 28, and Theorem 27. To clarify, $\tilde{\theta}$ is a parameter configuration that exists, and we use it only for our theoretical arguments, not $\tilde{\theta}$ in our algorithm. Rather, $\hat{\theta}$ is obtained and used algorithmically.

Chapter 3

Continuous-Time Training Dynamics

3.1 Gradient flow as a model for stochastic gradient descent

Consider a neural network $f_\theta \colon \mathcal{X} \to \mathbb{R}$ with parameter $\theta \in \mathbb{R}^p$, data $X_1, \dots, X_N \in \mathcal{X}$, and $Y_1, \dots, Y_N \in \mathcal{Y} = \mathbb{R}$. Let $\mathcal{L} \colon \mathbb{R}^p \to \mathbb{R}$ be the loss function of the form

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(X_i), Y_i).$$

Gradient descent (GD) updates the parameters with

$$\theta^{k+1} = \theta^k - \alpha \nabla \mathcal{L}(\theta^k)$$

for $k = 0, 1, \dots$, where $\theta^0 \in \mathbb{R}^p$ is a starting point and $\alpha > 0$ is the learning rate. In general, the learning rate α can depend on the iteration, but let us consider the constant stepsize case for the sake of simplicity.

Stochastic gradient descent (SGD) updates the parameters with

$$\begin{aligned} \theta^{k+1} &= \theta^k - \alpha \nabla_\theta \ell(f_\theta(X_{i(k)}), Y_{i(k)})\big|_{\theta=\theta^k} \\ &= \theta^k - \alpha(\nabla \mathcal{L}(\theta^k) + \varepsilon^k) \end{aligned}$$

for $k = 0, 1, \dots$, where $i(k) \sim \text{Uniform}(\{1, \dots, N\})$ is an IID sequence of random indices. Here, $\varepsilon^k = \nabla_\theta \ell(f_\theta(X_{i(k)}), Y_{i(k)})\big|_{\theta=\theta^k} - \nabla \mathcal{L}(\theta^k)$ is a martingale difference sequence, i.e.,

$$\mathbb{E}\left[\varepsilon^k \,|\, \theta^k\right] = 0.$$

In deep learning theory, one often considers the simplified training dynamics through gradient flow (GF):

$$\dot{\theta}(t) = -\nabla\mathcal{L}(\theta(t)),$$

where $\theta(0) = \theta^0$.

In what sense does the continuous-time deterministic process (GF) approximate the the discrete-time stochastic algorithm (SGD) well? The differential form of the intuition is that for GD,

$$\dot{\theta}(t) \approx \frac{\theta^{k+1} - \theta^k}{\alpha} = -\nabla\mathcal{L}(\theta^k),$$

where $t = \alpha k$. The idea is that we view α as the discretization step, so the total ellapsed "time" is the number of iterations k times the discretization size α. The integral form of the intuition is

$$\theta_k = \theta_0 - \alpha\sum_{i=0}^{k}\nabla\mathcal{L}(\theta^i) \approx \theta^0 - \int_0^t \nabla\mathcal{L}(\theta(s))\,ds.$$

For SGD, the integral form is more easily interpretable:

$$\theta_k = \theta_0 - \alpha\sum_{i=0}^{k}\nabla\mathcal{L}(\theta^i) + \underbrace{\alpha\sum_{i=0}^{k}\varepsilon^i}_{\mathcal{O}(\sqrt{k}\alpha)=\mathcal{O}(\sqrt{t\alpha})} \approx \theta^0 - \int_0^T \nabla\mathcal{L}(\theta(d))\,ds.$$

To bound the noise term, we assume $\mathbb{E}[\|\varepsilon^k\|^2 \,|\, \theta_k] < \tau^2 < \infty$ and argue that

$$\mathrm{Var}\left(\alpha\sum_{i=0}^{k}\varepsilon^i\right) \le k\alpha^2\tau^2 = \alpha t\tau^2 \to 0$$

as $\alpha \to 0$, fixed t, and $k = \lfloor t/\alpha \rfloor$.

Theorem 29. *[GD→GF] Assume $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ is differentiable. Assume $\nabla\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}^p$ is L-Lipschitz continuous and M-bounded.*[1] *For $\alpha > 0$, let $\{\theta^k_{(\alpha)}\}_{k=0,1,\ldots}$ be the GD sequence generated with learning rate α starting from θ^0, i.e.,*

$$\theta^{k+1}_{(\alpha)} = \theta^k_{(\alpha)} - \alpha\nabla\mathcal{L}(\theta^k_{(\alpha)})$$

[1]This assumption is meant to simplify the analysis, but, in general, it is sufficient to exclude cases where $\nabla\mathcal{L}$ varies too rapidly, such as $\theta^2\sin(1/\theta)$. One can establish the same result under the assumption $\nabla\mathcal{L}(\theta) \le C_1 + C_2 B(\|\theta\|)$, where $C_1 > 0$, $C_2 > 0$, and $B\colon \mathbb{R}_+ \to \mathbb{R}$ is an increasing function.

with $\theta^0_{(\alpha)} = \theta^0$. *Let* $\theta(t)$ *be the gradient flow starting from* $\theta(0) = \theta^0$. *Then for any* $T < \infty$,

$$\sup_{t\in[0,T]} \|\theta(t) - \theta^{\lfloor t/\alpha \rfloor}_{(\alpha)}\| \to 0$$

as $\alpha \to 0$.

Proof. For notational simplicity, we drop the subscript and write θ^k for $\theta^k_{(\alpha)}$. For $k \in N$,

$$\mathcal{E}_k = \theta(k\alpha) - \theta^k$$

to denote the error between GF and GD. Note, $\mathcal{E}_0 = 0$. Then,

$$\begin{aligned}
\mathcal{E}_{k+1} &= \theta((k+1)\alpha) - \theta^{k+1} \\
&= \theta(k\alpha) - \int_0^\alpha \nabla\mathcal{L}(\theta(k\alpha + s))\, ds - \theta^k + \alpha\nabla\mathcal{L}(\theta^k) \\
&= \mathcal{E}_k - \int_0^\alpha \nabla\mathcal{L}(\theta(k\alpha + s)) - \nabla\mathcal{L}(\theta(k\alpha))\, ds - \alpha\left(\nabla\mathcal{L}(\theta(k\alpha)) - \nabla\mathcal{L}(\theta^k)\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
\|\mathcal{E}_{k+1}\| &\le \|\mathcal{E}_k\| + \alpha\|\nabla\mathcal{L}(\theta(k\alpha)) - \nabla\mathcal{L}(\theta^k))\| + \int_0^\alpha \|\nabla\mathcal{L}(\theta(k\alpha + s)) - \nabla\mathcal{L}(\theta(k\alpha))\|\, ds \\
&\le \|\mathcal{E}_k\| + \alpha L\|\theta(k\alpha) - \theta^k\| + L\int_0^\alpha \|\theta(k\alpha + s) - \theta(k\alpha)\|\, ds \\
&= (1 + L\alpha)\|\mathcal{E}_k\| + L\int_0^\alpha \left\|\int_0^s \nabla\mathcal{L}(\theta(k\alpha + t))dt\right\|\, ds \\
&\le (1 + L\alpha)\|\mathcal{E}_k\| + \frac{ML\alpha^2}{2}.
\end{aligned}$$

Setting $C = 1 + L\alpha$ and $P = \frac{ML\alpha^2}{2}$ gives,

$$\begin{aligned}
\|\mathcal{E}_0\| &= 0 \\
\|\mathcal{E}_1\| &\le P \\
\|\mathcal{E}_2\| &\le (1+C)P \\
&\vdots \\
\|\mathcal{E}_k\| &\le (1 + C + \cdots + C^{k-1})P \\
&= \frac{C^k - 1}{C-1} P \\
&= \frac{M\alpha}{2}((1+L\alpha)^k - 1) \\
&\le \frac{M\alpha}{2}(e^{L\alpha k} - 1) \\
&\le \frac{(e^{Lt} - 1)M}{2}\alpha,
\end{aligned}$$

where $t = \alpha k$. Therefore, for any $T < \infty$,

$$\sup_{t\in[0,T]} \|\mathcal{E}_{\lfloor t/\alpha \rfloor}\| = \sup_{t\in[0,T]} \|\theta(\lfloor t/\alpha \rfloor \alpha) - \theta^{\lfloor t/\alpha \rfloor}\| \to 0$$

as $\alpha \to 0$. Since

$$\|\theta(t) - \theta(\lfloor t/\alpha \rfloor \alpha)\| = \|\int_{\lfloor t/\alpha \rfloor \alpha}^{t} \dot{\theta}(s) ds\| \le M|t - \lfloor t/\alpha \rfloor \alpha| \le M\alpha,$$

we conclude

$$\sup_{t\in[0,T]} \|\theta(t) - \theta^{\lfloor t/\alpha \rfloor}\| \le \sup_{t\in[0,T]} \|\mathcal{E}_{\lfloor t/\alpha \rfloor}\| + \sup_{t\in[0,T]} \|\theta(t) - \theta(\lfloor t/\alpha \rfloor \alpha)\| \to 0 \text{ as } \alpha \to 0.$$

as $\alpha \to 0$. □

Theorem 30 (SGD≈GD)**.** *Assume $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ is differentiable. Assume $\nabla\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}^p$ is L-Lipschitz continuous and M-bounded. For $\alpha > 0$, let $\{\phi^k_{(\alpha)}\}_{k=0,1,\ldots}$ be the SGD sequence generated with learning rate α starting from ϕ^0, i.e.,*

$$\phi^{k+1}_{(\alpha)} = \phi^k_{(\alpha)} - \alpha(\nabla\mathcal{L}(\phi^k_{(\alpha)}) + \epsilon^k)$$

with $\phi^0_{(\alpha)} = \phi^0$. Assume ε^k satisfies $\mathbb{E}[\varepsilon^k \mid \phi^k_{(\alpha)}] = 0$ and $\mathbb{E}[\|\varepsilon^k\|^2 \mid \phi^k] \le \tau^2 < \infty$. For $\alpha > 0$, let $\{\theta^k_{(\alpha)}\}_{k=0,1,\ldots}$ be the GD sequence generated with learning rate α starting from $\theta^0 = \phi^0$. Then for any $T < \infty$, we have

$$\sup_{t\in[0,T]} \mathbb{E}\|\phi^{\lfloor t/\alpha \rfloor}_{(\alpha)} - \theta^{\lfloor t/\alpha \rfloor}_{(\alpha)}\|^2 \to 0$$

as $\alpha \to 0$.

Proof. For notational simplicity, we drop the subscript and write ϕ^k for $\phi^k_{(\alpha)}$. Denote the (deterministic) gradient descent iterates as

$$\theta^{k+1} = \theta^k - \alpha\nabla(\theta^k),$$

with $\theta^0 = \phi^0$. By Theorem 29, we have

$$\sup_{t\in[0,T]} \|\theta(t) - \theta^{\lfloor t/\alpha \rfloor}\| \to 0$$

as $\alpha \to 0$. Thus, it is enough to show that $\sup_{t\in[0,T]} \|\phi^{\lfloor t/\alpha \rfloor} - \theta^{\lfloor t/\alpha \rfloor}\| \to 0$ as $\alpha \to 0$ with probability 1. For $k \in N$, define

$$\mathcal{E}_k = \mathbb{E}\left[\|\theta^k - \phi^k\|^2\right].$$

Note, $\mathcal{E}_0 = 0$. Then,

$$\begin{aligned}
\mathbb{E}\left[\|\theta^{k+1} - \phi^{k+1}\|^2 \,|\, \phi^k\right] &= \|\theta^k - \phi^k\|^2 - 2\alpha\langle\theta^k - \phi^k, \mathbb{E}[\nabla L(\theta^k) - \nabla L(\phi^k) - \varepsilon^k \,|\, \phi^k]\rangle \\
&\qquad + \alpha^2\mathbb{E}[\|\nabla L(\theta^k) - \nabla L(\phi^k) - \varepsilon^k\|^2 \,|\, \phi^k] \\
&= \|\theta^k - \phi^k\|^2 - 2\alpha\langle\theta^k - \phi^k, \nabla L(\theta^k) - \nabla L(\phi^k)\rangle \\
&\qquad + \alpha^2(\|\nabla L(\theta^k) - \nabla L(\phi^k)\|^2 + \tau^2) \\
&\leq \|\theta^k - \phi^k\|^2 + 2\alpha\|\theta^k - \phi^k\|\|\nabla L(\theta^k) - \nabla L(\phi^k)\| \\
&\qquad + \alpha^2\|\nabla L(\theta^k) - \nabla L(\phi^k)\|^2 + \alpha^2\tau^2 \\
&= (1 + \alpha L)^2\|\theta^k - \phi^k\|^2 + \alpha^2\tau^2.
\end{aligned}$$

By taking the full expectation and using the tower property of total expectation, we have

$$\mathcal{E}_{k+1} \leq (1 + \alpha L)^2\mathcal{E}_k + \alpha^2\tau^2.$$

With an inductive reasoning similar to that of Theorem 29, we get

$$\begin{aligned}
\mathcal{E}_k &\leq \frac{(1 + \alpha L)^{2k} - 1}{(1 + \alpha L)^2 - 1}\alpha^2\tau^2 = \frac{((1 + \alpha L)^{2k} - 1}{(\alpha L^2 + 2L)}\alpha\tau^2 \\
&\leq \frac{(e^{2k\alpha L} - 1)\tau^2}{\alpha L^2 + 2L}\alpha = \frac{(e^{2tL} - 1)\tau^2}{\alpha L^2 + 2L}\alpha \to 0
\end{aligned}$$

as $\alpha \to 0$. □

Combining

Corollary 2 (SGD→GF)**.** *Consider the setup of Theorem 30. For any $T < \infty$, we have*

$$\sup_{t\in[0,T]} \mathbb{E}\|\phi(t) - \phi^{\lfloor t/\alpha \rfloor}_{(\alpha)}\|^2 \to 0$$

as $\alpha \to 0$.

Proof. This immediately follows from combining Theorems 29 and 30. □

Corollary 3. *Consider the setup of Theorem 30. Further, assume* $\|\varepsilon^k\|^2 \leq \tau^2$ *almost surely. For any* $T < \infty$*, we have*

$$\sup_{t\in[0,T]} \|\phi(t) - \phi_{(\alpha)}^{\lfloor t/\alpha \rfloor}\|^2 \to 0$$

as $\alpha \to 0$ *in probability.*

Proof outline. This requires a probabilistic argument. Let $N \in \mathbb{N}$ and $\varepsilon > 0$. $\delta > 0$?? Let

$$\mathcal{E}(t) = \|\phi(t) - \phi_{(\alpha)}^{\lfloor t/\alpha \rfloor}\|^2$$

Let $t_i = Ti/N$ for $i = 1, \ldots, N$. Then with probability $1 - \delta$, we have

$$\mathcal{E}(t_i) < \varepsilon/2$$

for all $i = 1, \ldots, N$. By the boundedness assumptions, we have

$$|\mathcal{E}(t_i) - \mathcal{E}(t_i + s)| < \varepsilon/2$$

for all $s \in (0, T/N)$. Therefore,

$$\sup_{t\in[0,T]} \mathcal{E}(t) < \varepsilon.$$

□

However, this approximation removes all dependence on noise. Therefore, one can consider the update

$$\theta_{(\alpha)}^{k+1} = \theta_{(\alpha)}^{k} - \alpha \nabla \mathcal{L}(\theta_{(\alpha)}^{k}) + \sqrt{\alpha}\varepsilon^k.$$

In integral form, we have

$$\begin{aligned}\theta_{(\alpha)}^{k} &= \theta^0 - \alpha \sum_{i=0}^{k-1} \nabla \mathcal{L}(\theta_{(\alpha)}^{i}) + \sqrt{\alpha} \sum_{i=0}^{k-1} \varepsilon^i \\ &\approx \theta^0 - \int_0^{\lfloor t/\alpha \rfloor} \mathcal{L}(\theta(s))\, ds + \sqrt{2} B(t),\end{aligned}$$

where $B(t)$ is a Brownian motion. The argument follows from reasoning similar to Donsker's theorem. In this case, under suitable assumptions, the process $\{\theta_{(\alpha)}^{\lfloor t/\alpha \rfloor}\}_{t\geq 0}$ converges in distribution to the solution of the stochastic differential equation

$$d\theta = -\nabla \mathcal{L}(\theta) dt + \sqrt{2} dB(t)$$

This follows from Donsker's theorem. However, the SDE analysis is beyond the scope of this course, so we do not pursue it.

3.2 Continuous-time analysis of gradient flow

We have established that gradient flow approximates SGD. Now let us analyze the continuous-time dynamics of gradient flow

$$\dot{\theta} = -\nabla \mathcal{L}(\theta), \qquad \theta(0) = \theta_0.$$

In general, the training loss of finite deep neural networks is non-convex. However, the loss becomes convex in the infinite-width (NTK) limit, so we study the convex setup.

Theorem 31. *Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be a differentiable convex function. Assume a minimizer $\theta_\star$ exists. Then the solution to gradient flow $\{\theta(t)\}_{t\geq 0}$ exhibits the rate*

$$\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_\star) \leq \frac{\|\theta(0) - \theta^\star\|^2}{2t}.$$

Proof. Consider the energy function (also called a Lyapunov function)

$$\mathcal{E}(t) = t(\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)) + \frac{1}{2}\|\theta - \theta^\star\|^2.$$

Then

$$\begin{aligned}
\dot{\mathcal{E}}(t) &= (\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)) + t\langle \nabla \mathcal{L}(\theta), \dot{\theta}\rangle + \langle \theta - \theta^\star, \dot{\theta}\rangle \\
&= \mathcal{L}(\theta) - \mathcal{L}(\theta^\star) + \langle \theta^\star - \theta, \nabla \mathcal{L}(\theta)\rangle - t\|\nabla \mathcal{L}(\theta)\|^2 \\
&\leq -t\|\nabla \mathcal{L}(\theta)\|^2 \\
&\leq 0,
\end{aligned}$$

where the first inequality follows from convexity of $\mathcal{L}$. Since $\dot{\mathcal{E}} \leq 0$, we conclude

$$t(\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)) \leq t(\mathcal{L}(\theta) - \mathcal{L}(\theta^\star)) + \frac{1}{2}\|\theta(t) - \theta^\star\|^2 = \mathcal{E}(t) \leq \mathcal{E}(0) = \frac{1}{2}\|\theta^0 - \theta^\star\|^2.$$

□

The training loss in deep learning are usually non-convex and non-differentiable (due to ReLU activation functions). The "gradient flow" dynamics of non-differentiable functions can be formalized using the notion of subgradients, but we will not pursue that direction in this course. We will instead consider the dynamics of non-convex differentiable losses.

Theorem 32. *Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be continuously differentiable. Assume $\inf_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta) > -\infty$. If a solution to the gradient flow ODE exists, then*

$$\int_0^\infty \|\nabla \mathcal{L}(\theta(t))\|^2 dt < \infty.$$

So if $\theta(t)$ converges, then $\nabla\mathcal{L}(\theta(t)) \to 0$.

Proof. Consider the energy function

$$\mathcal{E}(t) = \int_0^t \|\nabla \mathcal{L}(\theta(s))\|^2 \, ds + \mathcal{L}(\theta(t)).$$

Then from $\dot\theta = -\nabla\mathcal{L}(\theta)$,

$$\frac{d}{dt}\mathcal{E}(t) = \|\nabla\mathcal{L}(\theta)\|^2 + \left\langle \nabla\mathcal{L}(\theta), \dot\theta \right\rangle = 0.$$

Thus, $\mathcal{E}(t)$ is constant as a function of time and $\mathcal{L}(\theta(t))$ is monotonically nonincreasing. Finally,

$$\begin{aligned}\int_0^\infty \|\nabla \mathcal{L}(\theta(t))\|^2 dt &= \mathcal{L}(\theta(0)) - \lim_{s\to\infty} \mathcal{L}(\theta(s)) \\ &\le \mathcal{L}(\theta(0)) - \inf_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta) < \infty.\end{aligned}$$

□

3.3 Second-order dynamics as a model for SGD with momentum

The gradient descent with momentum updates the iterates with

$$\theta_{k+1} = \theta_k + \beta(\theta_k - \theta_{k-1}) - \alpha\nabla\mathcal{L}(\theta_k),$$

where $\beta > 0$ is the momentum coefficient, α is the stepsize, and $\theta_0 = \theta_{-1}$ are the starting points. We can equivalently describe the algorithm as

$$\begin{aligned}v^{k+1} &= \beta v^k - \sqrt{\alpha}\nabla\mathcal{L}(\theta_k) \\ \theta^{k+1} &= \theta^k + \sqrt{\alpha} v^k,\end{aligned}$$

with initial condition $\theta^0 \in \mathbb{R}^d$ and $v^0 = 0$. The equivalence can be established by basic induction.

Let $\beta = 1 - \gamma\sqrt{\alpha}$. Then

$$\frac{v^{k+1} - v^k}{\sqrt{\alpha}} = -\gamma v^k - \nabla\mathcal{L}(\theta_k)$$
$$\frac{\theta^{k+1} - \theta^k}{\sqrt{\alpha}} = v^k$$

and we obtain the limiting ODE:

$$\dot{v}(t) = -\gamma v(t) - \nabla\mathcal{L}(\theta(t))$$
$$\dot{\theta}(t) = v(t),$$

with $\theta(0) = \theta_0$, $v(0) = 0$, and $t = \sqrt{\alpha}k$. Equivalently, we can express this as a second-order ODE:

$$\ddot{\theta}(t) + \gamma\dot{\theta}(t) + \nabla\mathcal{L}(\theta(t)) = 0$$

with $\theta(0) = \theta_0$ and $\dot{\theta}(0) = 0$.

This second-order ODE has a physical interpretation. If there is a particle with mass 1 subject to potential $\mathcal{L}$ and friction with friction coefficient $\gamma > 0$, then the dynamics of such a particle is governed by the same ODE.

Theorem 33. *Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be continuously differentiable. Assume $\inf_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta) > -\infty$. If a solution to the momentum gradient flow ODE exists, then*

$$\int_0^\infty \|\dot{\theta}(t)\|^2 dt < \infty.$$

So if $\dot{\theta}(t) \to 0$, then $\nabla\mathcal{L}(\theta(t)) \to 0$.

Proof. Then we have

$$\mathcal{E} = \frac{1}{2}\|\dot{\theta}\|^2 + \mathcal{L}(\theta) + \gamma\int_0^t \|\dot{\theta}\|^2\, ds$$

is a conserved quantity. Since $\mathcal{L}(\theta) > \inf_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta)$, we have

$$\int_0^\infty \|\dot{\theta}(t)\|^2 dt < \infty.$$

If $\theta(t)$ converges, then $\dot{\theta}(t) \to 0$ and $\ddot{\theta}(t) \to 0$. Then $\nabla\mathcal{L}(\theta) = 0$. □

Theorem 34. *Assume $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ is differentiable. Assume $\nabla\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}^p$ is L-Lipschitz continuous and M-bounded. Let $\{\theta^k_{(\alpha)}\}_{k=0,1,\cdots}$ be the GD sequence generated with momentum $\beta = 1-\gamma\sqrt{\alpha}$ and step size α starting from θ^0, i.e.,*

$$\theta^{k+1}_{(\alpha)} = \theta^k_{(\alpha)} + \beta(\theta^k_{(\alpha)} - \theta^{k-1}_{(\alpha)}) - s\nabla\mathcal{L}(\theta^k_{(\alpha)}),$$

with $\theta^0_{(\alpha)} = \theta^{-1}_{(\alpha)} = \theta^0$. Let $\theta(t)$ be the solution of momentum ODE starting from $\theta(0) = \theta^0$ and $\dot\theta(0) = 0$. Then for any $T < \infty$,

$$\sup_{t\in[0,T]} \|\theta(t) - \theta^{\lfloor t/\sqrt{s}\rfloor}_{(\alpha)}\| \to 0$$

as $s \to 0$.

Proof. For notational simplicity, we drop the subscript and write θ^k for $\theta^k_{(\alpha)}$.

$$\begin{aligned} v^{k+1} &= \beta v^k - \sqrt{\alpha}\nabla\mathcal{L}(\theta_k)\\ \theta^{k+1} &= \theta^k + \sqrt{\alpha}v^k, \end{aligned}$$

The ODE can be written as

$$\begin{aligned} \dot v(t) &= -\gamma v(t) - \nabla\mathcal{L}(\theta(t))\\ \dot\theta(t) &= v(t), \end{aligned}$$

Lemma 19. *Under said assumptions,* $\sup_{t\in[0,T]} \max\{\|v(t)\|, \|\dot v(t)\|\} = C < \infty$.

For $k = 0, 1, \ldots,$ denote

$$A_k = \|\theta^k - \theta(k\sqrt{\alpha})\|, \qquad B_k = \|v^k - v(k\sqrt{\alpha})\|$$

whose initial values are $A_0 = B_0 = 0$. First, we have

$$\begin{aligned} A_{k+1} &= \left\|\theta^{k+1} - \theta\left((k+1)s\right)\right\| \\ &= \left\|\theta^k + \sqrt{\alpha}v^k - \theta((k\sqrt{\alpha})) - \int_0^{\sqrt{\alpha}} \dot\theta(k\sqrt{\alpha}+t)\, dt\right\| \\ &\le \left\|\theta^k - \theta(k\sqrt{\alpha})\right\| + \left\|\sqrt{\alpha}v^k - \sqrt{\alpha}v(k\sqrt{\alpha})\right\| + \left\|\int_0^{\sqrt{\alpha}} v(k\sqrt{\alpha}+t) - v(k\sqrt{\alpha})\, dt\right\| \\ &\le A_k + \sqrt{\alpha}B_k + \frac{1}{2}C\alpha, \end{aligned}$$

where the last inequality follows from

$$\left\| \int_0^{\sqrt{\alpha}} v(k\sqrt{\alpha}+t) - v(k\sqrt{\alpha})\, dt \right\| \le \int_0^{\sqrt{\alpha}} \int_0^t \|\dot{v}(k\sqrt{\alpha}+v)\|\, dv dt \le \frac{1}{2} C\alpha.$$

Next, we have

$$\begin{aligned}
B_{k+1} &= \|v^{k+1} - v((k+1)\sqrt{\alpha})\| \\
&= \left\| \beta v^k - \sqrt{\alpha}\nabla\mathcal{L}(\theta^{k+1}) - v(k\sqrt{\alpha}) - \int_0^{\sqrt{\alpha}} \dot{v}(k\sqrt{\alpha}+t)\, dt \right\| \\
&= \Bigg\| (1-\gamma\sqrt{\alpha})(v^k - v(k\sqrt{\alpha})) \\
&\quad - \sqrt{\alpha}\nabla\mathcal{L}(\theta^{k+1}) + \sqrt{\alpha}\nabla\mathcal{L}(\theta^k) - \sqrt{\alpha}\nabla\mathcal{L}(\theta^k) + \sqrt{\alpha}\nabla\mathcal{L}(\theta(k\sqrt{\alpha})) \\
&\quad + \gamma \int_0^{\sqrt{\alpha}} v(k\sqrt{\alpha}+t) - v(k\sqrt{\alpha})\, dt + \int_0^{\sqrt{\alpha}} \nabla\mathcal{L}(\theta(k\sqrt{\alpha}+t)) - \nabla\mathcal{L}(\theta(k\sqrt{\alpha}))\, dt \Bigg\| \\
&\le (1-\gamma\sqrt{\alpha})B_k + L\alpha\|v^k\| + \sqrt{\alpha}LA_k + \gamma\frac{1}{2}C\alpha + C\frac{1}{2}L\alpha \\
&\le B_k + L\sqrt{\alpha}A_k + C'\alpha
\end{aligned}$$

for some constant $C' < \infty$. The last inequality follows from $\|v^k\| \le \|v^0\| + k\sqrt{\alpha}j \le TM$.

Let

$$U = \begin{bmatrix} 1 & \sqrt{\alpha} \\ L\sqrt{\alpha} & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} -1 & 1 \\ \sqrt{L} & \sqrt{L} \end{bmatrix}}_{=V} \underbrace{\begin{bmatrix} 1-\sqrt{L\alpha} & 0 \\ 0 & 1+\sqrt{L\alpha} \end{bmatrix}}_{=\mathrm{diag}(\lambda_1,\lambda_2)} \underbrace{\begin{bmatrix} -1/2 & 1/(2\sqrt{L}) \\ 1/2 & 1/(2\sqrt{L}) \end{bmatrix}}_{=V^{-1}}.$$

Then we have

$$\begin{bmatrix} A_{k+1} \\ B_{k+1} \end{bmatrix} \le U \begin{bmatrix} A_k \\ B_k \end{bmatrix} + \begin{bmatrix} 0 \\ C'\alpha \end{bmatrix}.$$

So we have

$$\begin{aligned}
\begin{bmatrix} A_k \\ B_k \end{bmatrix} &\leq (U^0 + U^1 + \cdots + U^{k-1}) \begin{bmatrix} 0 \\ C\alpha \end{bmatrix} \\
&= V \begin{bmatrix} \frac{1-(1-\sqrt{L\alpha})^k}{\sqrt{L\alpha}} & 0 \\ 0 & \frac{(1+\sqrt{L\alpha})^k - 1}{\sqrt{L\alpha}} \end{bmatrix} V^{-1} \begin{bmatrix} 0 \\ C'\alpha \end{bmatrix} \\
&= \frac{C'\alpha}{2\sqrt{\alpha L}} \begin{bmatrix} (1+\sqrt{L\alpha})^k + (1-\sqrt{L\alpha})^k - 2 \\ (1+\sqrt{L\alpha})^k - (1-\sqrt{L\alpha})^k \end{bmatrix} \\
&\leq \frac{C'\sqrt{\alpha}}{2\sqrt{L}} \begin{bmatrix} \exp(\sqrt{L\alpha}k) - 1 \\ \exp(\sqrt{L\alpha}k) \end{bmatrix} \\
&= \frac{C'\sqrt{\alpha}}{2\sqrt{L}} \begin{bmatrix} \exp(\sqrt{L}t - 1 \\ \exp(\sqrt{L}t) \end{bmatrix}.
\end{aligned}$$

This converges to 0 as $\alpha \to 0$.

□

Theorem 35. *Assume $\mathcal{L}$ is differentiable. Assume $\nabla\mathcal{L} : \mathbb{R}^p \to \mathbb{R}^p$ is L-Lipschitz continous and M-bounded. For $s > 0$, let $\{\theta^k_{(s)}\}_{k=0,1,\cdots}$ be the SGD sequence generated with monentum coefficient β and step size s starting from θ^0 and $\theta^1 = \theta^0 - s\nabla\mathcal{L}(\theta^0)$, i.e.,*

$$\theta^{k+1}_{(s)} = \theta^k_{(s)} + \beta(\theta^k_{(s)} - \theta^{k-1}_{(s)}) - s(\nabla\mathcal{L}(\theta^k_{(s)}) + \epsilon^k),$$

with $\theta^0_{(s)} = \theta^0$ and $\theta^1_{(s)} = \theta^1$.Assume ε^k satisfies $\mathbb{E}[\varepsilon^k \mid \theta^k_{(\beta)}] = 0$ and $\mathbb{E}[\|\varepsilon^k\|^2 \mid \theta^k] \leq \tau^2 < \infty$. Assume β satisfies the following condition as function of s:

$$1 - \beta = r\sqrt{s} + O(s) \text{ as } s \to 0$$

Let $\theta(t)$ be the solution of limiting ODE starting from $\theta(0) = \theta^0$ and $\dot{\theta}(0) = 0$. Then for any $T < \infty$,

$$\sup_{t\in[0,T]} \mathbb{E}\|\theta(t) - \theta^{\lfloor t/\sqrt{s}\rfloor}_{(s)}\|^2 \to 0$$

as $s \to 0$.

Proof. For notational simplicity, we drop the subscript and write θ^k for $\theta^k_{(s)}$. Denote the (deterministic) gradient descent iterates as

$$\phi^{k+1} = \phi^k + \beta(\phi^k - \phi^{k-1}) - s\nabla\mathcal{L}(\phi^k)$$

with $\phi^0 = \theta^0$. By Theorem 34, we have

$$\sup_{t\in[0,T]} \|\theta(t) - \phi^{\lfloor t/\beta \rfloor}\| \to 0$$

as $\beta \to 0$. Thus it is enought to show that $\sup_{t\in[0,T]} \mathbb{E}\|\theta^{\lfloor t/\beta \rfloor} - \phi^{\lfloor t/\beta \rfloor}\| \to 0$ as $\beta \to 0$. Let $\Phi^{k+1} = \frac{\phi^{k+1}-\phi^k}{s}$ and $\Theta^{k+1} = \frac{\theta^{k+1}-\theta^k}{s}$ for $k \in N_{\geq 0}$. Then iterates are equivalent to

$$\begin{aligned}\Phi^{k+1} &= \beta\Phi^k - \nabla\mathcal{L}(\phi^k)\\ \phi^{k+1} &= \phi^k + s\Phi^{k+1}\end{aligned}$$

and

$$\begin{aligned}\Theta^{k+1} &= \beta\Theta^k - \nabla\mathcal{L}(\theta^k) - \epsilon^k\\ \theta^{k+1} &= \theta^k + s\Theta^{k+1}\end{aligned}$$

with $\phi^0 = \theta^0$ and $\Phi^0 = \Theta^0 = 0$. Denote by $a_k = \sqrt{\mathbb{E}\|\phi^k - \theta^k\|^2}$ and $b_k = \sqrt{\mathbb{E}\|\Phi^k - \Theta^k\|^2}$, whose initial values are $a_0 = b_0 = 0$.
1) Using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}a_{k+1}^2 &= \mathbb{E}\|\phi^{k+1} - \theta^{k+1}\|^2\\ &= \mathbb{E}\|\phi^k - \theta^k\|^2 + 2|\mathbb{E}\langle\phi^k - \theta^k, s\Phi^{k+1} - s\Theta^{k+1}\rangle| + s^2\mathbb{E}\|\Phi^{k+1} - \Theta^{k+1}\|^2\\ &\leq a_k^2 + 2sa_kb_{k+1} + s^2b_{k+1}^2 = (a_k + 2sb_{k+1})^2.\end{aligned}$$

Hence, $a_k \leq a_{k-1} + b_k \leq \cdots \leq sS_k$ where $S_k = b_0 + b_1 + \cdots + b_k$.

2) Since ϵ^k is independent with ϕ^k, θ^k, Φ^k and Θ^k, we can get

$$\begin{aligned}b_{k+1}^2 &= \mathbb{E}\|\Phi^{k+1} - \Theta^{k+1}\|^2\\ &\leq \mathbb{E}\|\beta\Phi^k - \beta\Theta^k\|^2 + 2|\mathbb{E}\langle\beta\Phi^k - \beta\Theta^k, \nabla\mathcal{L}(\phi^k) - \nabla\mathcal{L}(\theta^k) - \epsilon^k\rangle|\\ &\quad + \mathbb{E}\|\nabla\mathcal{L}(\phi^k) - \nabla\mathcal{L}(\theta^k) - \epsilon^k\|^2\\ &\overset{(i)}{\leq} \beta^2b_k^2 + 2\beta Lb_ka_k + L^2a_k^2 + \tau^2\\ &\overset{(ii)}{\leq} (b_k + La_k)^2 + \tau^2\\ &\leq (b_k + LsS_k)^2 + \tau^2\end{aligned}$$

Inequality (i) follows from L-Lipschitz continuity of $\nabla\mathcal{L}$ and Cauchy-Schwarz inequality. Inequality (ii) follows from $\beta \leq 1$.

For bounding a_k, we define sequence η_k as

$$c_{k+1}^2 = (c_k + LsD_k)^2 + \tau^2$$

where $D_k = c_0 + c_1 + \cdots + c_k$ with $c_0 = 0$. Then, $b_k \leq c_k$ and $c_k \leq c_{k+1}$ is obvious. Since c_k is increasing sequence, we get

$$c_{k+1}^2 = (c_k + LsD_k)^2 + \tau^2 \leq (c_k + Lskc_k)^2 + \tau^2 = (Lsk+1)^2 c_k^2 + \tau^2.$$

Due to $k\sqrt{s} \leq T$, bound is changed to $c_{k+1}^2 \leq (LT\sqrt{s}+1)^2 c_k^2 + \tau^2$. By induction on k, it holds that

$$\begin{aligned} c_k^2 &\leq \tau^2 + (1 + TL\sqrt{s})^2\tau^2 + \cdots + (1 + TL\sqrt{s})^{2(k-1)}\tau^2 \\ &= \frac{P^k - 1}{P - 1}\tau^2 \end{aligned}$$

, denoting by $P := (1 + TL\sqrt{s})^2$. Hence, with Cauchy-Schwarz inequality, we obtain

$$D_k \leq \sqrt{k(c_0^2 + \cdots c_k^2)} \leq \sqrt{k\left(\frac{P^{k+1} - 1}{(P-1)^2} - \frac{k+1}{P-1}\right)}$$

Thus,

$$a_k \leq sS_k \leq sD_k \leq \sqrt{\frac{(P^{k+1} - 1)ks^2}{(P-1)^2} - \frac{(k+1)ks^2}{P-1}}$$

Using following arguments, this yields $a_k = O(\sqrt{s})$ as $s \to 0$.

- $\frac{(P^{k+1}-1)ks^2}{(P-1)^2} = \frac{k\sqrt{s}((TL\sqrt{s}+1)^{2k+2}-1)}{(T^2L^2\sqrt{s}+2TL)^2}\sqrt{s} \to 0$ as $s \to 0$ for $k \leq \lfloor T/\sqrt{s} \rfloor$ ($\lim_{s\to 0}(TL\sqrt{s}+1)^{2k+2}$ has limit $e^{T^2L} < \infty$)
- $\frac{(k+1)ks^2}{P-1} = \frac{k\sqrt{s}(k\sqrt{s}+\sqrt{s})}{T^2L^2\sqrt{s}+2TL}\sqrt{s} \to 0$ as $s \to 0$.

□

Let $\mu > 0$. Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be differentiable. We say $\mathcal{L}$ is μ-strongly convex if

$$\mathcal{L}(\varphi) \geq \mathcal{L}(\theta) + \langle \nabla\mathcal{L}(\theta), \varphi - \theta\rangle + \frac{\mu}{2}\|\varphi - \theta\|^2, \qquad \forall\, \varphi, \theta \in \mathcal{L}.$$

Equivalently, $\mathcal{L}$ is μ-strongly convex if $\mathcal{L}(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex.

For μ-strongly convex functions, momentum gradient flow exhibits an accelerated rate over gradient flow without momentum.

Theorem 36. *Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be a differentiable μ-strongly convex function. If a solution to the gradient flow ODE (without momentum) exists, then*

$$\mathcal{L}(\theta(t)) - \mathcal{L}_\star \leq e^{-2\mu t}(\mathcal{L}(\theta(0)) - \mathcal{L}(\theta_\star)) = \mathcal{O}(e^{-2\mu t})$$

Proof outline. Consider

$$\mathcal{E}(t) = e^{2\mu t}\left(\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_\star)\right)$$

and use the inequality

$$\mathcal{L}(\theta) - \mathcal{L}(\theta_\star) \leq \frac{1}{2\mu}\|\nabla\mathcal{L}(\theta)\|^2$$

to show $\frac{d}{dt}\mathcal{E}(t) \leq 0$. □

Theorem 37. *Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be a differentiable μ-strongly convex function. Let $\gamma = 2\sqrt{\mu}$. If a solution to the momentum gradient flow ODE exists, then*

$$\begin{aligned}\mathcal{L}(\theta(t)) - \mathcal{L}_\star &\leq e^{-\sqrt{\mu}t}\left(\mathcal{L}(\theta(0)) - \mathcal{L}(\theta_\star) + \frac{1}{2}\left\|\dot\theta + \sqrt{\mu}(\theta - \theta_\star)\right\|^2\right)\\ &= \mathcal{O}(e^{-\sqrt{\mu}t})\end{aligned}$$

Proof outline. Consider

$$\mathcal{E}(t) = e^{\sqrt{\mu}t}\left(\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_\star) + \frac{1}{2}\left\|\dot\theta + \sqrt{\mu}(\theta - \theta_\star)\right\|^2\right)$$

and show $\frac{d}{dt}\mathcal{E}(t) \leq 0$. □

Chapter 4

Gaussian process

Let $\mathcal{X}$ be a nonempty sample space. We say $\{f(x)\}_{x\in\mathcal{X}}$ is a *stochastic process* if $f\colon \mathcal{X} \to \mathbb{R}^d$ is a random function.[1] We say $\{f(x)\}_{x\in\mathcal{X}}$ is a *Gaussian process* (GP) if for any $N \in \mathbb{N}$ and $x_1, \ldots, x_N \in \mathcal{X}$, the joint marginal distribution

$$(f(x_1), \ldots, f(x_N)) \in \mathbb{R}^{Nd}$$

is a Gaussian distribution. The distribution of a Gaussian process is characterized fully by its mean and covariance.

$$\mu(x) = \mathbb{E}_f[f(x)], \qquad \Sigma(x, x') = \mathbb{E}_f[(f(x) - \mu(x))(f(x') - \mu(x'))^\intercal] \in \mathbb{R}^{d\times d}.$$

The covariance kernel $\Sigma(x, x')$ is necessarily a mvPDK.

Theorem 38. *Let $\mathcal{X}$ be a nonempty set. Let $\mu\colon \mathcal{X} \to \mathbb{R}^d$ be an arbitrary function and let $\Sigma\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{d\times d}$ be a mvPDK. Then there exists a probability space $(\mathcal{W}, \mathcal{F}, \mathbb{P})$ and $f(\cdot;\cdot)\colon \mathcal{X}\times\mathcal{W} \to \mathbb{R}^d$ such that $\{f(x; w)\}_{x\in\mathcal{X}}$ is a Gaussian peocess with mean function μ and covariance kernel Σ.*

The precise meaning of the "existence" of GPs may be difficult to grasp for those who are not already familiar with measure-theoretic probability theory. If you do not have the background, there is no need to worry. For our purposes, you can accept this existence result as a statement that GPs are mathematically well-defined.

[1] More precisely, let there be a probability space $(\mathcal{W}, \mathcal{F}, P)$, where $\mathcal{W}$ is is the sample space, $\mathcal{F}$ is the event space (the σ-algebra), and P is the probability measure. Then $f(\cdot;\cdot)\colon \mathcal{X} \times \mathcal{W} \to \mathbb{R}^d$, and given random outcome $w \in \mathcal{W}$, we view $f(\cdot; w)\colon \mathcal{X} \to \mathbb{R}^d$ as the instantiation of the random function. Alternatively, the sample space $\mathcal{W}$ is itself be a space of functions, such as $\mathcal{W} = \mathcal{C}(\mathcal{X})$, and $\mathcal{F}$ is a σ-algebra (the "cylindrical" σ-algebra), and P is a probability measure assigning probabilities on sets of realizations of the random function $f \in \mathcal{W}$.

Proof outline. The key step is to extend the probability space of the joint marginals using the Kolmogorov consistency theorem. □

We write $f \sim \mathcal{GP}(\mu, \Sigma)$ to denote that f is a Gaussian process with mean μ and covariance kernel Σ, i.e., for any $x \in \mathcal{X}$, we have

$$\mathbb{E}_{f\sim\mathcal{GP}(\mu,\Sigma)}[f(x)] = \mu(x), \qquad \mathbb{E}_{f\sim\mathcal{GP}(\mu,\Sigma)}[(f(x)-\mu(x))(f(x')-\mu(x'))^\intercal] = \Sigma(x, x').$$

4.1 Neural network Gaussian process

Let us characterize neural networks at initialization as Gaussian processes. Consider the depth-L multilayer perceptron

$$\begin{aligned}
f_\theta(x) &= y_L \\
y_L &= z_L, & z_L &= A_L y_{L-1} + b_L \in \mathbb{R}^{n_L} \\
y_{L-1} &= \sigma(z_{L-1}), & z_{L-1} &= A_{L-1} y_{L-2} + b_{L-1} \in \mathbb{R}^{n_{L-1}} \\
&\vdots \\
y_2 &= \sigma(z_2), & z_2 &= A_2 y_1 + b_2 \in \mathbb{R}^{n_2} \\
y_1 &= \sigma(z_1), & z_1 &= A_1 x + b_1 \in \mathbb{R}^{n_1}
\end{aligned}$$

where $x \in \mathbb{R}^{n_0}$, $A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, $b_\ell \in \mathbb{R}^{n_\ell}$, and $n_L = 1$. (To clarify, σ is applied element-wise.) Assume the parameters are initialized via the "LeCun initialization"

$$(A_\ell)_{ij} \sim \mathcal{N}(0, \sigma_A^2/n_{\ell-1}), \qquad (b_\ell)_i \sim \mathcal{N}(0, \sigma_b^2).$$

The input x is considered non-random. In the following, we establish that f_θ is a Gaussian process in the infinite-width limit at initialization. In this infinite limit, $n_1, \ldots, n_{L-1} \to \infty$, but n_0 and n_L remain finite.

The conclusion will be that $\{f_\theta(x)\}_{x\in\mathcal{X}}$ is a GP. To get there, we characterize the distribution of each "pre-activation value" $\{z_\ell(x)\}_{x\in\mathcal{X}}$ as GPs.

First layer. Note that

$$(z_1)_i(x) = (A_1)_{i,:}x + (b_1)_i = \sum_{j=1}^{n_0} (A_1)_{i,j}x_j + (b_1)_i$$

is a (finite) sum of independent zero-mean Gaussians and is therefore a zero-mean Gaussian. Moreover, the n_0 components

$$\{(z_1)_1(x)\}_{x\in\mathcal{X}}, \{(z_1)_2(x)\}_{x\in\mathcal{X}}, \ldots, \{(z_1)_{n_1}(x)\}_{x\in\mathcal{X}}$$

are independent in the sense that $(z_1)_i(x)$ and $(z_1)_j(x')$ are independent random variables for any $i \neq j$ and $x, x' \in \mathbb{R}^{n_0}$.

We now characterize the mean and covariance. The mean is zero:

$$\mathbb{E}[(z_1)_i(x)] = 0$$

for $i = 1, \ldots, n_1$. The different components are independent:

$$\mathbb{E}[(z_1)_i(x)(z_1)_j(x')] = 0$$

for $i \neq j$. The non-trivial correlations:

$$\begin{aligned}
\mathbb{E}[(z_1)_i(x)(z_1)_i(x')] &= \mathbb{E}[((b_1)_i)^2] + \mathbb{E}[((A_1)_{i,:}x)((A_1)_{i,:}x')] \\
&\quad + \text{cross terms zero since } A \text{ and } b \text{ are independent} \\
&= \sigma_b^2 + \mathbb{E}[\text{Trace}((A_1)_{i,:}x'x^\intercal(A_1)_{i,:}^\intercal)] \\
&= \sigma_b^2 + \mathbb{E}[\text{Trace}(x'x^\intercal(A_1)_{i,:}^\intercal(A_1)_{i,:})] \\
&= \sigma_b^2 + \text{Trace}(x'x^\intercal\mathbb{E}[(A_1)_{i,:}^\intercal(A_1)_{i,:}]) \\
&= \sigma_b^2 + \text{Trace}\left(x'x^\intercal\frac{\sigma_A^2}{n_0}I\right) \\
&= \sigma_b^2 + \frac{\sigma_A^2}{n_0}x^\intercal x' \\
&:= \Sigma^{(1)}(x, x')
\end{aligned}$$

for $i = 1, \ldots, n_1$. Note that $n_0 \nrightarrow \infty$. So far, there is no infinite-width argument yet.

From this analysis, we conclude that $\{(z_1)_1(x)\}_{x\in\mathcal{X}}, \ldots, \{(z_1)_{n_1}(x)\}_{x\in\mathcal{X}}$ are n_1 IID scalar-valued GPs and

$$(z_1)_i \sim \mathcal{GP}(0, \Sigma^{(1)})$$

for $i = 1, \ldots, n_1$. This also means

$$z_1 \sim \mathcal{GP}(0, \Sigma^{(1)} \otimes I_{n_1}),$$

where $I_{n_1} \in \mathbb{R}^{n_1\times n_1}$ is the $n_1 \times n_1$ identity matrix. To clarify, $\{z_1(x)\}_{x\in\mathcal{X}}$ is a $\mathbb{R}^{n_1}$-valued zero-mean GP and with covariance kernel

$$(\Sigma^{(1)} \otimes I_{n_1})(x, x') = \text{diag}(\Sigma^{(1)}(x, x'), \ldots, \Sigma^{(1)}(x, x')).$$

Intermediate layers. Note that

$$(z_2)_i(x) = (A_2)_{i,:}y_1 + (b_2)_i = \sum_{j=1}^{n_1}(A_2)_{i,j}(y_1)_j + (b_2)_i$$

is a sum of non-Gaussians. ($(y_1)_j$ is not a Gaussian, and even if it were a Gaussian, the product $(A_2)_{i,j}(y_1)_j$ would not be Gaussian.) We will later take the limit $n_1 \to \infty$ so that $(z_2)_i$ converges to a Gaussian. The n_2 components

$$\{(z_2)_1(x)\}_{x\in\mathcal{X}}, \{(z_2)_2(x)\}_{x\in\mathcal{X}}, \ldots, \{(z_2)_{n_2}(x)\}_{x\in\mathcal{X}}$$

are independent.

We now characterize the mean and covariance. The mean is zero:

$$\mathbb{E}[(z_2)_i(x)] = 0$$

for $i = 1, \ldots, n_1$. The different components are independent:

$$\mathbb{E}[(z_2)_i(x)(z_2)_j(x')] = 0.$$

for $i \neq j$. The non-trivial correlations:

$$\begin{aligned}
\mathbb{E}[(z_2)_i(x)(z_2)_i(x')] &= \sigma_b^2 + \mathbb{E}[((A_2)_{i,:}\sigma(z_1(x)))((A_2)_{i,:}\sigma(z_1(x')))] \\
&= \sigma_b^2 + \frac{\sigma_A^2}{n_1}\mathbb{E}[\sigma(z_1(x))^\intercal\sigma(z_1(x'))] \\
&= \sigma_b^2 + \frac{\sigma_A^2}{n_1}\sum_{k=1}^{n_1}\mathbb{E}\left[\sigma((z_1(x))_k)\sigma((z_1(x'))_k)\right] \\
&= \sigma_b^2 + \sigma_A^2\mathbb{E}_{f\sim\mathcal{GP}(0,\Sigma^{(1)})}[\sigma(f(x))\sigma(f(x'))] \\
&:= \Sigma^{(2)}(x,x')
\end{aligned}$$

for $i = 1, \ldots, n_1$. Finally, let $n_1 \to \infty$. Since

$$(z_2)_i(x) = \sum_{j=1}^{n_1}(A_2)_{i,j}(y_1)_j + (b_2)_i$$

is a sum of n_1 IID random variables, it converges to a Gaussian with covariance described by $\Sigma^{(2)}(x,x')$ by the central limit theorem.

From this analysis, we conclude that $\{(z_2)_1(x)\}_{x\in\mathcal{X}}, \ldots, \{(z_2)_{n_2}(x)\}_{x\in\mathcal{X}}$ are n_2 IID scalar-valued GPs and

$$(z_2)_i \sim \mathcal{GP}(0, \Sigma^{(2)})$$

for $i = 1, \ldots, n_2$. This also means

$$z_2 \sim \mathcal{GP}(0, \Sigma^{(2)} \otimes I_{n_2}).$$

Recursively, we have

$$\Sigma^{(\ell+1)}(x, x') = \sigma_b^2 + \sigma_A^2 \mathbb{E}_{f \sim \mathcal{GP}(0, \Sigma^{(\ell)})}[\sigma(f(x))\sigma(f(x'))]$$

for $\ell = 1, \ldots, L-1$. From this analysis, we conclude that $\{(z_\ell)_1(x)\}_{x \in \mathcal{X}}, \ldots, \{(z_\ell)_{n_\ell}(x)\}_{x \in \mathcal{X}}$ are n_ℓ IID scalar-valued GPs and

$$(z_\ell)_i \sim \mathcal{GP}(0, \Sigma^{(\ell)})$$

for $i = 1, \ldots, n_2$. This also means

$$z_\ell \sim \mathcal{GP}(0, \Sigma^{(\ell)} \otimes I_{n_\ell})$$

for $\ell = 2, \ldots, L$. In particular, we conclude

$$f_\theta \sim \mathcal{GP}(0, \Sigma^{(L)}).$$

Chapter 5

Neural tangent kernel

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty. Consider the setup with training data $X_1, \dots, X_N \in \mathcal{X}$ and corresponding labels $Y_1, \dots, Y_N \in \mathcal{Y}$. For the sake of concreteness and simplicity, let $\mathcal{Y} = \mathbb{R}^k$ and $Y_i = f_\star(X_i)$ for $i = 1, \dots, N$ for some true unknown $f_\star$.[1] Consider a neural network $f_\theta(x) \in \mathbb{R}^k$ that is continuous in both the input x and parameter θ and, furthermore, continuously-differentiable in the parameter θ in the sense that $\nabla_\theta f_\theta(x)$ is well defined for all θ and x and is continuous both in θ and x.[2] Let $P \in \mathcal{P}(\mathcal{X})$ be a probability measure that is compactly supported.[3] The primary example to consider is the empirical distribution

$$P_{\text{emp}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i},$$

where $X_1, \dots, X_N$ are the training data. Consider the *risk* $R\colon L^2(P; \mathbb{R}^k) \to \mathbb{R}$ defined as

$$R[f_\theta] = \mathbb{E}_{X \sim P}[\ell(f_\theta(X); f_\star(X))],$$

where $\ell\colon \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ is continuous.[4] Consider training through

$$\underset{\theta \in \mathbb{R}^P}{\text{minimize}} \quad R[f_\theta].$$

[1] The assumption that labels are deterministic is not necessary, and is merely made to simplify notation.

[2] These continuity assumptions can be relaxed, but it is a simple and realistic assumption that ensures the integrals and expectations we consider are well defined (measurable and integrable).

[3] Compact support is not necessary, but it is a simple assumption that ensures the integrals and expectations we consider are well defined (integrable).

[4] Continuity of ℓ is not necessary, but it is a simple and realistic assumption that ensures the integrals and expectations we consider are well defined (measurable and integrable).

To clarify, $L^2(P;\mathbb{R}^k)$ is the equivalence class of $\mathbb{R}^k$-valued square integrable functions with respect to P. For any $f, g \in L^2(P;\mathbb{R}^k)$, the associated inner product is

$$\langle f, g\rangle_{L^2(P;\mathbb{R}^k)} = \mathbb{E}_{X\sim P}[f(X)^\intercal g(X)].$$

(The fact that $L^2(P;\mathbb{R}^k)$ is a space of equivalence classes, rather than functions, will revisited later.) Let $\mathcal{C}(\mathcal{X};\mathbb{R}^k)$ be the vector space of continuous functions from $\mathcal{X}$ to $\mathbb{R}^k$. We do not equip $\mathcal{C}(\mathcal{X};\mathbb{R}^k)$ with a metric.[5] Note that $f_\theta \in \mathcal{C}(\mathcal{X};\mathbb{R}^k) \subset L^2(P;\mathbb{R}^k)$ (with some abuse of notation).

Notation. $i : \mathbb{R}^{\mathcal{X}} \to L^2(P;\mathbb{R}^k), i(f) =$ equivalence class containing $f \in L^2(P;\mathbb{R}^k)$
$i^+ : L^2(P;\mathbb{R}^k) \to \mathbb{R}^{\mathcal{X}}, i(f) =$ equal to f on $\mathcal{X}$, 0 otherwise
$L_{\Theta_t} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}, L_{\Theta_t}[f](x) = \int \Theta(x,x')f(x')dP(x')$
$L^{(1)}_{\Theta_t} : L^2(P;\mathbb{R}^k) \to \mathbb{R}^{\mathcal{X}}$, $L^{(1)}_{\Theta_t} = L_{\Theta_t} \circ i^+$
$L^{(2)}_{\Theta_t} : L^2(P;\mathbb{R}^k) \to L^2(P;\mathbb{R}^k)$, $L^{(2)}_{\Theta_t} = i \circ L_{\Theta_t} \circ i^+$
$L^{(3,x)}_{\Theta_t} : L^2(P;\mathbb{R}^k) \to \mathbb{R}^k$, $L^{(3,x)}_{\Theta_t} = L_x \circ L_{\Theta_t} \circ i^+$ where L_x is evaluation at x.

5.1 Kernel gradient flow via the chain rule

We analyze the training dynamics induced by gradient flow

$$\frac{d\theta(t)}{dt} = -\left(\nabla_\theta R[f_\theta]\right)\Big|_{\theta=\theta(t)}.$$

For notational conciseness, we often omit the time dependence and write

$$\frac{d\theta}{dt} = -\nabla_\theta R[f_\theta].$$

A key abstraction of the NTK theory is to analyze the dynamics of the prediction function $f_{\theta(t)}$, rather than the parameters $\theta(t)$. Translating the gradient flow dynamics of $\theta(t)$ to dynamics of the prediction function requires some chain-rule calculations.

To effectively understand these calculations, we carry it out three times. First, we will do so formally (non-rigorously), proceeding as if we can just apply the chain rule of vector calculus. Second, we rigorously and carefully define the relevant notion of derivatives and properly carry out the derivation. Third, we carry out the derivation in a simpler, more concrete, instance and observe agreement.

[5] However, we will consider pointwise limits when defining things like $\frac{d}{dt}f_{\theta(t)}$. So we are, in effect, implicitly consider the topology of point-wise convergence.

5.1.1 Formal calculations for gradient flow

We shall carry out formal calculations, proceeding as if we can just apply the chain rule:

$$\begin{aligned}\frac{d\theta}{dt} &= -\nabla_\theta R[f_\theta] && (P \times 1)\\ &= -(D_\theta R[f_\theta])^\intercal && (1 \times P)^\intercal\\ &= -\left(\frac{\partial f_\theta}{\partial \theta}\right)^\intercal \left(\frac{\partial R}{\partial f}\right)^\intercal && (P \times \dim(f))(\dim(f) \times 1).\end{aligned}$$

Further proceeding,

$$\begin{aligned}\frac{df_\theta}{dt} &= \frac{\partial f_\theta}{\partial \theta}\frac{d\theta}{dt} && (\dim(f) \times P)(P \times 1)\\ &= -\underbrace{\frac{\partial f_\theta}{\partial \theta}\left(\frac{\partial f_\theta}{\partial \theta}\right)^\intercal}_{=\Theta_t}\left(\frac{\partial R}{\partial f}\right)^\intercal && (\dim(f) \times P)(P \times \dim(f))(\dim(f) \times 1)\\ &= -\Theta_t\left(\frac{\partial R}{\partial f}\right)^\intercal && (\dim(f) \times \dim(f))(\dim(f) \times 1).\end{aligned}$$

We call

$$\frac{df_{\theta(t)}}{dt} = -\Theta_t\left(\frac{\partial R}{\partial f}\right)^\intercal = -\Theta_t \nabla_f R$$

kernel gradient flow in contrast to (regular) gradient flow in the function space

$$\frac{d}{dt}f_t = -\left(\frac{\partial R}{\partial f}\right)^\intercal = \nabla_f R.$$

The (regular) gradient flow often has better convergence properties than kernel gradient flow, as discussed in Section 5.1.5, but it does not model reality. In practical deep learning, we update θ through SGD, and this induces an update on f_θ. Kernel gradient flow models this. However, regular gradient flow directly updates the prediction function f, but there is no practical mechanism for directly updating the prediction function.

As another aside, kernel gradient flow can be considered a gradient flow with the descent direction preconditioned by Θ_t. The preconditioning takes the gradient $\nabla_f R$ into an appropriate tangent space, as discussed in Section 5.1.4.

5.1.2 Rigorous derivation of kernel gradient flow

Now let us carry out the same derivation rigorously. Assume the risk $R\colon L^2(P;\mathbb{R}^k) \to \mathbb{R}$ is (Fréchet) differentiable at any point $f_0 \in L^2(P;\mathbb{R}^k)$ with derivative

$$\partial_f R|_{f_0} \in L^2(P;\mathbb{R}^k),$$

which is also called the functional derivative of R at f_0. Specifically, $\partial_f R|_{f_0}$ is the derivative of R at f_0 defined via

$$R[f_0 + \delta] = R[f_0] + \langle \partial_f R|_{f_0}, \delta \rangle_{L^2(P;\mathbb{R}^k)} + o(\|\delta\|_{L^2(P;\mathbb{R}^k)})$$

for small $\delta \in L^2(P;\mathbb{R}^k)$. For simplicity, we often write $\partial_f R = \partial_f R|_{f_0}$.

We apply the chain rule on training via gradient flow:

$$\begin{aligned}
\frac{d\theta_p}{dt} &= -\frac{d}{d\theta_p} R[f_\theta] \\
&= -\left\langle \frac{\partial f_\theta}{\partial \theta_p}, \partial_f R \right\rangle_{L^2(P;\mathbb{R}^k)}
\end{aligned}$$

for $p = 1, \ldots, P$. This chain rule requires some additional assumptions, but it does hold when P has finite support. (Cf. Homework problem.) For any $x \in \mathcal{X}$, we again apply the chain rule to get

$$\begin{aligned}
\frac{d}{dt} f_\theta(x) &= \frac{\partial f_\theta(x)}{\partial \theta} \frac{d\theta}{dt} \\
&= \sum_{p=1}^{P} \frac{\partial f_\theta(x)}{\partial \theta_p} \frac{d\theta_p}{dt} \\
&= -\sum_{p=1}^{P} \frac{\partial f_\theta(x)}{\partial \theta_p} \left\langle \frac{\partial f_\theta}{\partial \theta_p}, \partial_f R \right\rangle_{L^2(P;\mathbb{R}^k)} \\
&= -\sum_{p=1}^{P} \frac{\partial f_\theta(x)}{\partial \theta_p} \mathbb{E}_{x' \sim P} \left[\left(\frac{\partial f_\theta(x')}{\partial \theta_p} \right)^\intercal \partial_f R(x') \right] \\
&= -\mathbb{E}_{x' \sim P} \left[\sum_{p=1}^{P} \frac{\partial f_\theta(x)}{\partial \theta_p} \left(\frac{\partial f_\theta(x')}{\partial \theta_p} \right)^\intercal \partial_f R(x') \right] \\
&= -\mathbb{E}_{x' \sim P} \left[\Theta_t(x, x') \partial_f R(x') \right] \\
&= -L_{\Theta_t}[\partial_f R](x).
\end{aligned}$$

To clarify, $\theta = \theta(t)$, i.e., θ is time-dependent, but we suppress the time-dependence to simplify the notation. Here, we define the *neural tangent kernel* (NTK) as

$$\begin{aligned}
\Theta_t(x, x') &= \sum_{p=1}^{P} \left(\frac{\partial f_\theta(x)}{\partial \theta_p} \right) \left(\frac{\partial f_\theta(x')}{\partial \theta_p} \right)^\intercal \\
&= \left(\frac{\partial f_\theta(x)}{\partial \theta} \right) \left(\frac{\partial f_\theta(x')}{\partial \theta} \right)^\intercal
\end{aligned}$$

or equivalently

$$\Theta_t = \sum_{p=1}^{P} \frac{\partial f_\theta}{\partial \theta_p} \otimes \frac{\partial f_\theta}{\partial \theta_p}.$$

In conclusion, we have the *kernel gradient flow*

$$\frac{d}{dt} f_\theta = -L_{\Theta_t}[\partial_f R],$$

which contrast with the regular gradient flow

$$\frac{d}{dt} f_t = -\partial_f R.$$

5.1.3 Special case: Quadratic function, empirical risk

For the sake of concretentess, consider the empirical risk minimization with quadratic loss

$$R_{\text{emp}}[f_\theta] = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \| f_\theta(X_i) - f^\star(X_i) \|^2.$$

Then, gradient flow on the parameter is

$$\begin{aligned} \frac{d\theta}{dt} &= -\frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \frac{1}{2} \| f_\theta(X_i) - f^\star(X_i) \|^2 \\ &= -\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial f_\theta(X_i)}{\partial \theta} \right)^\intercal (f_\theta(X_i) - f^\star(X_i)). \end{aligned}$$

Kernel gradient flow becomes

$$\begin{aligned} \frac{d}{dt} f_\theta(x) &= \frac{\partial f_\theta(x)}{\partial \theta} \frac{d\theta}{dt} \\ &= -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial f_\theta(x)}{\partial \theta} \left(\frac{\partial f_\theta(X_i)}{\partial \theta} \right)^\intercal (f_\theta(X_i) - f^\star(X_i)) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \Theta_t(x, X_i)(f_\theta(X_i) - f^\star(X_i)) \\ &= -\mathbb{E}_{x' \sim P_{\text{emp}}} \left[\Theta_t(x, x')(f_\theta(x') - f^\star(x')) \right] \\ &= -L_{\Theta_t}[f_\theta - f^\star](x). \end{aligned}$$

In conclusion, kernel gradient flow is

$$\frac{d}{dt} f_\theta = -L_{\Theta_t}[f_\theta - f^\star].$$

5.1.4 Tangent space interpretation

In the space of function $\mathcal{C}(\mathcal{X};\mathbb{R}^k)$, the space of configurations of the neural network f_θ

$$\mathcal{M} = \{f_\theta \,|\, \theta \in \mathbb{R}^P\} \subset \mathcal{C}(\mathcal{X};\mathbb{R}^k)$$

can be viewed as a something like a manifold of dimension P. For $f_{\theta_0} \in \mathcal{M}$, the tangent space is

$$\mathcal{T}_{f_{\theta_0}} = \left\{ \frac{d}{dt} f_{\theta(t)}|_{t=0} \,\middle|\, \theta(t) \text{ such that } \theta(0) = \theta_0 \right\}.$$

Since

$$\frac{d}{dt} f_{\theta(t)} = \frac{\partial f}{\partial \theta} \frac{d\theta}{dt}$$

and since $\frac{d\theta}{dt}|_{t=0} \in \mathbb{R}^P$ can be arbitrary, we have

$$\mathcal{T}_{f_{\theta_0}} = \text{span} \left\{ \frac{\partial f}{\partial \theta_p}\Big|_{\theta=\theta_0} \right\}_{p=1}^{P} = \left\{ \frac{\partial f}{\partial \theta}\Big|_{\theta=\theta_0} v \,\middle|\, v \in \mathbb{R}^P \right\}.$$

The linear map L_Θ defined by the neural tangent kernel at θ_0

$$\Theta(x, x') = \left(\frac{\partial f_\theta}{\partial \theta}(x)\Big|_{\theta=\theta_0} \right) \left(\frac{\partial f_\theta}{\partial \theta}(x')\Big|_{\theta=\theta_0} \right)^\intercal$$

is a mapping into the tangent space in the sense that $L_\Theta[h] \in \mathcal{T}_{f_{\theta_0}}$ for all $h \in \mathcal{C}(\mathcal{X};\mathbb{R}^k)$. The regular gradient flow

$$\frac{d}{dt} f_t = -\partial_f R$$

generates a trajectory that escapes the manifold $\mathcal{M}$. Therefore, the dynamics of gradient flow represents is not something we can mimic or realize using our neural network f_θ. However, kernel gradient flow

$$\frac{d}{dt} f_{\theta(t)} = -L_{\Theta_t}[\partial_f R]$$

generates a trajectory that stays within the $\mathcal{M}$, because L_{Θ_t} maps the functional derivative $\partial_f R$ onto the tangent space. In hindsight, however, that $f_{\theta(t)}$ stays within $\mathcal{M}$ is not surprising, since $f_{\theta(t)}$ is simply f_θ with $\theta = \theta(t)$ plugged in.

5.1.5 Convergence properties of kernel gradient flow

For regular gradient flow

$$\frac{d}{dt} f_t = -\partial_f R$$

if R is convex and if a minimizer $f_\star$ exists, then we have global convergence of the risk value

$$R[f_t] - R[f_\star] \leq \frac{\|f_0 - f_\star\|^2}{2t}$$

by arguments identical to that of Theorem 31. The assumption that R is convex is realistic; many commonly used loss functions such as the mean-squared error loss or the cross-entropy loss are indeed convex. The existence of a minimizer is arguably a mild assumption that holds in most cases.

However, kernel gradient flow

$$\frac{d}{dt} f_{\theta(t)} = -L_{\Theta_t}[\partial_f R],$$

in general, does not enjoy the same convergence properties. The risk does monotonically decrease, since

$$\begin{aligned}
\frac{d}{dt} R[f_{\theta(t)}] &= \left\langle \partial_f R, \frac{d}{dt} f_{\theta(t)} \right\rangle_{L^2(P;\mathbb{R}^k)} \\
&= -\mathbb{E}_{x\sim P}\left[(\partial_f R)(x)^\intercal (L_{\Theta_t}[\partial_f R])(x)\right] \\
&= -\mathbb{E}_{x,x'\sim P}\left[(\partial_f R)(x)^\intercal \Theta_t(x,x')(\partial_f R)(x')\right] \\
&\leq 0.
\end{aligned}$$

However, we cannot draw any conclusion regarding global convergence, i.e., it is possible that $\lim_{t\to\infty} R[f_{\theta(t)}] > R[f_\star]$.

This is to be expected, since minimizing $R[f_\theta]$ with respect to θ is non-convex optimization, which is an NP-hard problem class.

Since kernel gradient flow is not a real algorithm, it is not inconceivable that a continuous-time flow converges to global minima of non-convex functions. If so, the continuous-time process would require exponentially many steps to approximate with a discrete, implementable algorithm unless P=NP. (The overdamped Langevin equation, an SDE model of SGD, "converges" to the global minima but the convergence takes exponentially long.) In any case, kernel gradient flow does not converge to the global minimum, so one should not expect the discrete counterparts, GD and SGD, to do so either.

At a surface level, the roadblock in the analysis is that the kernel Θ_t is time-dependent; the time-dependence of $\theta(t)$ causes Θ_t to change (or "twist") as a function of time. A remarkable discovery of the NTK theory is that, in

the infinite-width limit, the kernel becomes time-independent, i.e., $\Theta_t = \Theta$ as width$\to \infty$. In this case, the kernel gradient flow

$$\frac{d}{dt} f_t = -L_\Theta[\partial_f R]$$

does converge to the global minimum with rate

$$R[f_t] - R[f_\star] \leq \mathcal{O}(1/t)$$

by arguments identical to that of the homework problem.

5.2 NTK at initialization

Consider the depth-L multilayer perceptron (MLP)

$$\begin{aligned}
f_\theta(x) &= y_L \\
y_L &= z_L, & z_L &= \frac{\sigma_A}{\sqrt{n_{L-1}}} A_L y_{L-1} + \sigma_b b_L \in \mathbb{R}^{n_L}, \\
y_{L-1} &= \sigma(z_{L-1}), & z_{L-1} &= \frac{\sigma_A}{\sqrt{n_{L-2}}} A_{L-1} y_{L-2} + \sigma_b b_{L-1} \in \mathbb{R}^{n_{L-1}}, \\
&\vdots \\
y_2 &= \sigma(z_2), & z_2 &= \frac{\sigma_A}{\sqrt{n_1}} A_2 y_1 + \sigma_b b_2 \in \mathbb{R}^{n_2}, \\
y_1 &= \sigma(z_1), & z_1 &= \frac{\sigma_A}{\sqrt{n_0}} A_1 x + \sigma_b b_1 \in \mathbb{R}^{n_1},
\end{aligned}$$

where $\sigma_A > 0$, $\sigma_b > 0$, $x \in \mathbb{R}^{n_0}$, $A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, and $b_\ell \in \mathbb{R}^{n_\ell}$. (We do not assume $n_L = 1$.) Assume the parameters are initialized as

$$(A_\ell)_{ij} \sim \mathcal{N}(0,1), \qquad (b_\ell)_i \sim \mathcal{N}(0,1).$$

This initialization is "equivalent" to the LeCun initialization we had see in the prior discussion of NNGPs.

Theorem 39. *In the limit $n_1, \ldots, n_{L-1} \to \infty$,*

$$f_\theta \sim \mathcal{GP}(0, \Sigma^{(L)} \otimes I_{n_L}),$$

where

$$\Sigma^{(1)}(x, x') = \frac{\sigma_A^2}{n_0} x^\intercal x' + \sigma_b^2$$

and

$$\Sigma^{(\ell+1)}(x, x') = \sigma_A^2 \mathbb{E}_{f \sim \mathcal{GP}(0, \Sigma^{(\ell)})}[\sigma(f(x))\sigma(f(x'))] + \sigma_b^2$$

for $\ell = 1, \ldots, L-1$.

However, the scaling is different. The limiting GP is identical at *at initialization*, but the scaling will alter the training dynamics of gradient flow. In a homework assignment, you will be see what happens when we use the usual scaling.

Since $f_\theta = z_L$ outputs a length n_L vector, its covariance is expressed by a matrix-valued ($\mathbb{R}^{n_L \times n_L}$-valued) PDK. (The covariance between $f_\theta(x)$ and $f_\theta(x')$ always defines a mvPDK, even for finite values of $n_1, \dots, n_{L-1}$.) The limiting covariance kernel, however, is a tensor product of a scalar-valued PDK and the $n_L \times n_L$ identity matrix $\Sigma^{(L)} \otimes I_{n_L}$, i.e.,

$$\operatorname{cov}(f_\theta(x), f_\theta(x')) \to \Sigma^{(L)}(x, x') I_{n_L}.$$

This means that, in the infinite-width limit, (i) $(f_\theta(x))_i$ and $(f_\theta(x))_j$ are uncorrelated and hence independent for $i \neq j$ and (ii) $(f_\theta(x))_1, \dots, (f_\theta(x))_{n_L}$ are IID scalar-valued zero-mean Gaussian processes with identical scalar-valued covariance kernel $\Sigma^{(L)}$.

For $\ell = 1, \dots, L$, define

$$\theta^{(\ell)} = (A_1, b_1, A_2, b_2, \dots, A_\ell, b_\ell)$$

to represent the neural network parameters up to and including layer ℓ. So, $\theta^{(L)} = \theta$. For $\ell = 1, \dots, L$, define

$$\Theta_t^{(\ell)}(x, x') = \left(\frac{\partial z_\ell(x)}{\partial \theta^{(\ell)}}\right) \left(\frac{\partial z_\ell(x')}{\partial \theta^{(\ell)}}\right)^\intercal,$$

So $\Theta_t = \Theta_t^{(L)}$.

Next, we analyze the training dynamics of the MLP in the infinite width-limit by characterizing the limiting kernel. First, we show that the NTK has a very nice limit at initialization, i.e., at time $t = 0$.

Theorem 40. *Assume $\sigma \colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous. In the limit $n_1, \dots, n_{L-1} \to \infty$,*

$$\Theta_0^{(L)} \to \Theta_0^{(L),\infty} \otimes I_{n_L}$$

in probability. Furthermore,

$$\Theta_0^{(1),\infty} = \Sigma^{(1)}$$

and

$$\Theta_0^{(\ell+1),\infty} = \Theta_0^{(\ell),\infty} \dot{\Sigma}^{(\ell+1)} + \Sigma^{(\ell+1)}, \qquad \ell = 1, \dots, L-1,$$

where $\Sigma^{(\ell)}$ is as defined in Theorem 39 and

$$\dot{\Sigma}^{(\ell+1)}(x, x') = \sigma_A^2 \mathbb{E}_{f \sim \mathcal{GP}(0, \Sigma^{(\ell)})}[\sigma'(f(x))\sigma'(f(x'))], \qquad \ell = 1, \dots, L-1.$$

To clarify, σ' denotes the derivative of σ, and $\Theta_0^{(\ell),\infty}\dot{\Sigma}^{(\ell+1)}$ denotes the pointwise product of two scalar-valued PDKs, so

$$\Theta_0^{(\ell+1),\infty}(x,x') = \Theta_0^{(\ell),\infty}(x,x')\dot{\Sigma}^{(\ell+1)}(x,x') + \Sigma^{(\ell+1)}(x,x').$$

Proof of Theorem 40. Throughout the proof, we set $t = 0$ and drop the time dependence to simplify the notation. We establish the claim via induction on the depth $\ell = 1, \ldots, L$. For the sake of simplicity, we will consider taking the limits

$$(n_1 \to \infty), (n_2 \to \infty), \ldots, (n_{L-1} \to \infty)$$

sequentially, i.e., we consider the limit

$$\lim_{n_L \to \infty} \lim_{n_{L-1} \to \infty} \cdots \lim_{n_2 \to \infty} \lim_{n_1 \to \infty} \boxed{}.$$

The claimed result holds more generally under the limit $n_1, \ldots, n_{L-1} \to \infty$, i.e., under the limit

$$\lim_{\min\{n_1,\ldots,n_{L-1}\} \to \infty} \boxed{},$$

but we will not present the argument here.

First, we have

$$\begin{aligned}
\Theta_{kk'}^{(1)}(x,x') &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \frac{\partial(z_1)_k(x)}{\partial(A_1)_{ij}}\frac{\partial(z_1)_{k'}(x')}{\partial(A_1)_{ij}} + \sum_{i=1}^{n_1} \frac{\partial(z_1)_k(x)}{\partial(b_1)_i}\frac{\partial(z_1)_{k'}(x')}{\partial(b_1)_j} \\
&= \frac{\sigma_A^2}{n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} x_j x'_j \delta_{ik}\delta_{ik'} + \sigma_b^2 \sum_{i=1}^{n_1} \delta_{ik}\delta_{ik'} \\
&= \frac{\sigma_A^2}{n_0} x^\intercal x' \delta_{kk'} + \sigma_b^2 \delta_{kk'} \\
&= \Sigma^{(1)}(x,x')\delta_{kk'}.
\end{aligned}$$

So, $\Theta^{(1)} = \Sigma^{(1)} \otimes I_{n_1}$. (This holds without $n_1 \to \infty$.)

Next, consider the limit $n_1, \ldots, n_{\ell-1} \to \infty$. We have

$$\begin{aligned}
\Theta_{kk'}^{(\ell+1)}(x,x') = &\sum_{i=1}^{n_{\ell+1}}\sum_{j=1}^{n_\ell} \frac{\partial(z_{\ell+1})_k(x)}{\partial(A_{\ell+1})_{ij}}\frac{\partial(z_{\ell+1})_{k'}(x')}{\partial(A_{\ell+1})_{ij}} + \sum_{i=1}^{n_{\ell+1}} \frac{\partial(z_{\ell+1})_k(x)}{\partial(b_{\ell+1})_i}\frac{\partial(z_{\ell+1})_{k'}(x')}{\partial(b_{\ell+1})_i} \\
&+ \left(\frac{\partial(z_{\ell+1})_k(x)}{\partial\theta^{(\ell)}}\right)\left(\frac{\partial(z_{\ell+1})_{k'}(x')}{\partial\theta^{(\ell)}}\right)^\intercal.
\end{aligned}$$

For the first two components, we have

$$\begin{aligned}
&\sum_{i=1}^{n_{\ell+1}}\sum_{j=1}^{n_\ell}\frac{\partial(z_{\ell+1})_k(x)}{\partial(A_{\ell+1})_{ij}}\frac{\partial(z_{\ell+1})_{k'}(x')}{\partial(A_{\ell+1})_{ij}}+\sum_{i=1}^{n_{\ell+1}}\frac{\partial(z_{\ell+1})_k(x)}{\partial(b_{\ell+1})_i}\frac{\partial(z_{\ell+1})_{k'}(x')}{\partial(b_{\ell+1})_i}\\
&=\frac{\sigma_A^2}{n_\ell}\sum_{i=1}^{n_{\ell+1}}\sum_{j=1}^{n_\ell}\sigma((z_\ell)_j(x))\sigma((z_\ell)_j(x'))\delta_{ik}\delta_{ik'}+\sigma_b^2\sum_{i=1}^{n_{\ell+1}}\delta_{ik}\delta_{ik'}\\
&=\frac{\sigma_A^2}{n_\ell}\sum_{j=1}^{n_\ell}\sigma((z_\ell)_j(x))\sigma((z_\ell)_j(x'))\delta_{kk'}+\sigma_b^2\delta_{kk'}\\
&\to\sigma_A^2\mathbb{E}_{f\sim\mathcal{GP}(0,\Sigma^{(\ell)})}[\sigma(f(x))\sigma(f(x'))]\delta_{kk'}+\sigma_b^2\delta_{kk'}\\
&=\Sigma^{(\ell)}(x,x')\delta_{kk'}.
\end{aligned}$$

For the third component, we have

$$\begin{aligned}
\frac{\partial(z_{\ell+1})_k(x)}{\partial\theta^{(\ell)}}&=\frac{\partial}{\partial\theta^{(\ell)}}\left(\frac{\sigma_A}{\sqrt{n_\ell}}(A_{\ell+1})_{k,:}\sigma(z_\ell(x))+\sigma_b(b_{\ell+1})_k\right)\\
&=\frac{\sigma_A}{\sqrt{n_\ell}}(A_{\ell+1})_{k,:}\operatorname{diag}(\sigma'(z_\ell(x)))\frac{\partial z_\ell(x)}{\partial\theta^{(\ell)}}\\
&=\frac{\sigma_A}{\sqrt{n_\ell}}\sum_{j=1}^{n_\ell}(A_{\ell+1})_{kj}\sigma'((z_\ell(x))_j)\frac{\partial(z_\ell)_j(x)}{\partial\theta^{(\ell)}}
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\left(\frac{\partial(z_{\ell+1})_k(x)}{\partial\theta^{(\ell)}}\right)\left(\frac{\partial(z_{\ell+1})_{k'}(x')}{\partial\theta^{(\ell)}}\right)^\intercal\\
&=\frac{\sigma_A^2}{n_\ell}\sum_{i=1}^{n_\ell}\sum_{j=1}^{n_\ell}(A_{\ell+1})_{ki}(A_{\ell+1})_{k'j}\sigma'((z_\ell(x))_i)\sigma'((z_\ell(x'))_j)\left(\frac{\partial(z_\ell)_i(x)}{\partial\theta^{(\ell)}}\right)\left(\frac{\partial(z_\ell)_j(x')}{\partial\theta^{(\ell)}}\right)^\intercal\\
&\to\frac{\sigma_A^2}{n_\ell}\sum_{i=1}^{n_\ell}\sum_{j=1}^{n_\ell}(A_{\ell+1})_{ki}(A_{\ell+1})_{k'j}\sigma'((z_\ell(x))_i)\sigma'((z_\ell(x'))_j)\Theta^{(\ell),\infty}(x,x')\delta_{ij}\\
&=\frac{\sigma_A^2}{n_\ell}\Theta^{(\ell),\infty}(x,x')\sum_{i=1}^{n_\ell}(A_{\ell+1})_{ki}(A_{\ell+1})_{k'i}\sigma'((z_\ell(x))_i)\sigma'((z_\ell(x'))_i)\\
&\to\Theta^{(\ell),\infty}(x,x')\sigma_A^2\mathbb{E}_{f\sim\mathcal{GP}(0,\Sigma^{(\ell)})}[\sigma'(f(x))\sigma'(f(x'))]\delta_{kk'}\\
&=\Theta^{(\ell),\infty}(x,x')\dot\Sigma^{(\ell+1)}(x,x')\delta_{kk'}.
\end{aligned}$$

So $\Theta^{(\ell+1)}\to(\Theta^{(\ell),\infty}\dot\Sigma^{(\ell+1)}+\Sigma^{(\ell+1)})\otimes I_{n_\ell}$.

□

Since $f_\theta = z_L$ outputs a length n_L vector, its kernel $\Theta_0^{(L)}$ is indeed matrix-valued ($\mathbb{R}^{n_L \times n_L}$-valued). The limiting kernel $\Theta_0^{(L)} \to \Theta_0^{(L),\infty} \otimes I_{n_L}$, however, has the simpler structure of a tensor product of a scalar-valued PDK and the identity matrix. The diagonal structure of $\Theta_0^{(L)}$ implies that

$$\frac{d}{dt} f_\theta = -L_{\Theta_0^{(L)}} \left[\partial R \big|_{f_\theta} \right]$$

splits into

$$\frac{d}{dt} (f_\theta)_i = -L_{\Theta_0^{(L),\infty}} \left[(\partial R)_i \big|_{f_\theta} \right], \qquad i = 1, \ldots, n_L.$$

To clarify, $\partial R \colon \mathcal{X} \to \mathbb{R}^{n_L}$ and $(\partial R)_i \colon \mathcal{X} \to \mathbb{R}$ is the ith coordinate of the ∂R. To put it differently, $(\partial R)_i \in L^2(P; \mathbb{R})$ and $(\partial R)_i(x) = e_i^\intercal \partial R(x)$, where $e_i \in \mathbb{R}^{n_L}$ is the ith unit vector, for $i = 1, \ldots, n_L$. So the gradient flow dynamics at initialization (and for time $t > 0$ as we later establish) split into nearly independent dynamics, still coupled through all coordinates of f_θ affecting $(\partial R)_i$. If, furthermore, the risk splits across the output coordinates, i.e., if

$$R[f] = \sum_{i=1}^{n_L} R_i[f_i]$$

then we have

$$\frac{d}{dt} (f_\theta)_i = -L_{\Theta_0^{(L),\infty}} \left[\partial (R_i) \big|_{(f_\theta)_i} \right], \qquad i = 1, \ldots, n_L$$

and the gradient flow dynamics completely splits.

One interpretation of the diagonality of the kernel is as follows. Note that

$$R[f_0 + \delta \otimes e_i] = R[f_0] + \langle (\partial_f R|_{f_0})_i, \delta \rangle_{L^2(P; \mathbb{R}^k)} + o(\|\delta\|)$$

for small $\delta \in L^2(P; \mathbb{R})$, where $e_i \in \mathbb{R}^{n_L}$ is the unit vector. To clarify,

$$(f_0 + \delta \otimes e_i)(x) = \begin{bmatrix} (f_0(x))_1 \\ (f_0(x))_2 \\ \vdots \\ (f_0(x))_{i-1} \\ (f_0(x))_i + \delta(x) \\ (f_0(x))_{i+1} \\ \vdots \\ (f_0(x))_{n_L} \end{bmatrix}.$$

In other words, $(\partial_f R|_{f_0})_i$ is the derivative of R with respect the infinitestimal changes in the ith output of the input function f_0. Therefore, the kernel gradient flow splitting as

$$\frac{d}{dt}(f_\theta)_i = -L_{\Theta_0^{(L),\infty}}\left[(\partial R)_i\big|_{f_\theta}\right], \qquad i = 1, \dots, n_L$$

means the derivative direction in the ith output of f_θ only affects the ith output of f_θ. (For finite neural networks, the kernel is not diagonal, so $(\partial_f R|_{f_0})_i$ affects the parameter θ, which in turn affects all of $(f_\theta)_1, \dots, (f_\theta)_{n_L}$.

5.3 Some preliminaries

We define the following order in probability notation. Let $X_1, X_2, \dots$ be a sequence of scalar-valued random variables and $a_1, a_2, \dots \in \mathbb{R}$ a sequence of deterministic scalars. We say

$$X_n = \mathcal{O}_p(a_n)$$

if for any $\varepsilon > 0$, there exists an $M > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}(|X_n/a_n| > M) < \varepsilon.$$

If

$$X_n = \mathcal{O}_p(1)$$

then we say $X_1, X_2, \dots$ is *stochastically bounded*. If $X_1, X_2, \dots$ is stochastically bounded and $\lim_{n\to\infty} b_n = \infty$, then

$$X_n/b_n \to 0$$

in probability.

Let $L\colon V \to W$ be a linear operator mapping from a normed vector space V to a normed vector space W. (Assume V is non-trivial, i.e., contains a nonzero element.) Then the operator norm of L is defined as

$$\|L\|_{\text{op}} = \sup_{v:\|v\|_V=1} \|L(v)\|_W.$$

If L is a bounded (continuous) linear operator, then $\|L\|_{\text{op}} < \infty$. In this case, we have

$$\|L(v)\|_W \le \|L\|_{\text{op}}\|v\|_V, \qquad \forall\, v \in V.$$

Lemma 20 (Grönwall's lemma)**.** *Let $T > 0$ and let $\mathcal{E}\colon [0,T] \to \mathbb{R}$ be differentiable on $(0,T)$. Let $\beta\colon [0,T] \to \mathbb{R}$. Assume*

$$\mathcal{E}'(t) \leq \beta(t)\mathcal{E}(t)$$

for all $t \in (0,T)$. Then,

$$\mathcal{E}(t) \leq \mathcal{E}(0) \exp\left(\int_0^t \beta(t)\ dt\right)$$

for all $t \in [0,T]$.

5.4 Invariance of NTK

Theorem 41 (NTK invariance)**.** *Assume $\sigma\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous and has bounded second derivative. Let $T > 0$ be fixed and consider the limit $n_1, \ldots, n_{L-1} \to \infty$. Assume*

$$\int_0^T \left\| \partial_f R|_{f_{\theta(t)}} \right\|_{L^2(P;\mathbb{R}^{n_L})}\ dt$$

is stochastically bounded. Then

$$\Theta_t^{(L)} \to \Theta^{(L),\infty} \otimes I_{n_L}$$

in probability, uniformly in $t \in [0,T]$ and pointwise for inputs (x, x').

Note that the right-hand side of the limit is time-independent. As a corollary, we expect the functional dynamics in the infinite-limit follow

$$\frac{d}{dt} f_\theta = -L_{\Theta^{(L,\infty)} \otimes I_{n_L}}[\partial_f R].$$

The proof of this result requires a somewhat arduous recursion argument. Therefore, for the sake of simplicity, we shall prove the result in the simplified 2-layer setup.

Proof of Theorem 41 for the 2-layer setup. Consider the depth-2 MLP

$$\begin{aligned} f_\theta(x) &= y_2 \\ y_2 &= z_2, \qquad & z_2 &= \frac{\sigma_A}{\sqrt{n_1}} A_2 y_1 + \sigma_b b_2 \in \mathbb{R}^{n_2}, \\ y_1 &= \sigma(z_1), \qquad & z_1 &= \frac{\sigma_A}{\sqrt{n_0}} A_1 x + \sigma_b b_1 \in \mathbb{R}^{n_1}, \end{aligned}$$

where $\sigma_A > 0$, $\sigma_b > 0$, $x \in \mathbb{R}^{n_0}$, $A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, and $b_\ell \in \mathbb{R}^{n_\ell}$

From the proof of Theorem 40, we have

$$\Theta_t^{(1)} = \Sigma^{(1)} \otimes I_{n_1}$$

for all t (even before taking any infinite width limit). Notably, the right-hand side is time-independent.

Next, note that

$$(\Theta_t^{(2)}(x,x'))_{kk'} = \sum_{i=1}^{n_2}\sum_{j=1}^{n_1} \frac{\partial (z_2)_k(x)}{\partial (A_2)_{ij}} \frac{\partial (z_2)_{k'}(x')}{\partial (A_2)_{ij}} + \sum_{i=1}^{n_2} \frac{\partial (z_2)_k(x)}{\partial (b_2)_i} \frac{\partial (z_2)_{k'}(x')}{\partial (b_2)_i} + \left(\frac{\partial (z_2)_k(x)}{\partial \theta^{(1)}}\right)\left(\frac{\partial (z_2)_{k'}(x')}{\partial \theta^{(1)}}\right)^\intercal.$$

To clarify, A_1, b_1, A_2, b_2 are time-dependent, while x and x' are considered fixed inputs. Consequently, $z_2(x)$ and $z_1(x)$ are also time-dependent. For the first two components,

$$\begin{aligned}
&\sum_{i=1}^{n_2}\sum_{j=1}^{n_1} \frac{\partial (z_2)_k(x)}{\partial (A_2)_{ij}} \frac{\partial (z_2)_{k'}(x')}{\partial (A_2)_{ij}} + \sum_{i=1}^{n_2} \frac{\partial (z_2)_k(x)}{\partial (b_2)_i} \frac{\partial (z_2)_{k'}(x')}{\partial (b_2)_i} \\
&= \frac{\sigma_A^2}{n_1} \sum_{i=1}^{n_2}\sum_{j=1}^{n_1} \sigma((z_1)_j(x))\sigma((z_1)_j(x'))\delta_{ik}\delta_{ik'} + \sigma_b^2 \sum_{i=1}^{n_2} \delta_{ik}\delta_{ik'} \\
&= \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \sigma((z_1)_j(x))\sigma((z_1)_j(x'))\delta_{k'k} + \sigma_b^2 \delta_{kk'} \\
&\overset{?}{\to} \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \sigma((z_1)_j(x))\sigma((z_1)_j(x'))\Big|_{t=0} \delta_{k'k} + \sigma_b^2 \delta_{kk'} \\
&= \sigma_A^2 \mathbb{E}_{f\sim \mathcal{GP}(0,\Sigma^{(1)})}[\sigma(f(x))\sigma(f(x'))]\delta_{k'k} + \sigma_b^2 \delta_{kk'}.
\end{aligned}$$

The limit in question, $\overset{?}{\to}$, holds if $(z_1)_j(x;t) \to (z_1)_j(x;0)$ at a rate uniform in

$j = 1, \dots, n_1$. For the third term,

$$
\begin{aligned}
&\left(\frac{\partial (z_2)_k(x)}{\partial \theta^{(1)}}\right)\left(\frac{\partial (z_2)_{k'}(x')}{\partial \theta^{(1)}}\right)^\intercal \\
&= \frac{\sigma_A^2}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}(A_2)_{ki}(A_2)_{k'j}\sigma'((z_1(x))_i)\sigma'((z_1(x'))_j)\left(\frac{\partial (z_1)_i(x)}{\partial \theta^{(1)}}\right)\left(\frac{\partial (z_1)_j(x')}{\partial \theta^{(1)}}\right)^\intercal \\
&= \frac{\sigma_A^2}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}(A_2)_{ki}(A_2)_{k'j}\sigma'((z_1(x))_i)\sigma'((z_1(x'))_j)\Theta_{ij}^{(1)}(x,x') \\
&= \Theta^{(1),\infty}(x,x')\frac{\sigma_A^2}{n_1}\sum_{j=1}^{n_1}(A_2)_{kj}(A_2)_{k'j}\sigma'((z_1(x))_j)\sigma'((z_1(x'))_j)
\end{aligned}
$$

So, to put it differently, we have

$$
\begin{aligned}
\left(\frac{\partial z_2(x)}{\partial \theta^{(1)}}\right)\left(\frac{\partial z_2(x')}{\partial \theta^{(1)}}\right)^\intercal &= \Theta^{(1),\infty}(x,x')\frac{\sigma_A^2}{n_1}\sum_{j=1}^{n_1}(A_2)_{:,j}\sigma'((z_1(x))_j)\sigma'((z_1(x'))_j)(A_2)_{:j}^\intercal \\
&\overset{?}{\to} \Theta^{(1),\infty}(x,x')\frac{\sigma_A^2}{n_1}\sum_{i=1}^{n_1}(A_2)_{:,j}\sigma'((z_1(x))_j)\sigma'((z_1(x'))_j)(A_2)_{:j}^\intercal\Big|_{t=0} \\
&= \Theta^{(1),\infty}(x,x')\sigma_A^2\dot{\Sigma}^{(2)}(x,x')I_{n_2}
\end{aligned}
$$

The limit in question, $\overset{?}{\to}$, holds if $(z_1)_j(x;t) \to (z_1)_j(x;0)$ and $(A_2)_{:j}(t) \to (A_2)_{:j}(0)$ at rates uniform in $j = 1, \dots, n_1$.

For notational simplicity, let us suppress the x dependency and write z_ℓ or $z_\ell(t)$, rather than $z_\ell(x;t)$. Since

$$
\begin{aligned}
\frac{d}{dt}(A_2)_{i,j}(t) &= -\left\langle \partial_f R|_{f_{\theta(t)}}, \frac{\partial z_2}{\partial (A_2)_{i,j}}\right\rangle_{L^2(P;\mathbb{R}^{n_2})} \\
&= -\frac{\sigma_A}{\sqrt{n_1}}\left\langle \partial_f R|_{f_{\theta(t)}}, \sigma((z_1)_j(\cdot;t))e_i\right\rangle_{L^2(P;\mathbb{R}^{n_2})} \\
&= -\frac{\sigma_A}{\sqrt{n_1}}\left\langle (\partial_f R|_{f_{\theta(t)}})_i, \sigma((z_1)_j(t))\right\rangle_{L^2(P;\mathbb{R})} \\
&= -\frac{\sigma_A}{\sqrt{n_1}}\left\langle (\partial_f R)_i, \sigma((z_1)_j)\right\rangle_{L^2(P;\mathbb{R})},
\end{aligned}
$$

where $e_i \in \mathbb{R}^{n_2}$ is the ith unit vector, and

$$\begin{aligned}
\frac{d}{dt}\left\|(A_2)_{:,j}(t) - (A_2)_{:,j}(0)\right\|_2 &\leq \left\|\frac{d}{dt}(A_2)_{:,j}(t)\right\|_2 \\
&= \frac{\sigma_A}{\sqrt{n_1}}\left\|\left(\langle(\partial_f R)_i, \sigma((z_1)_j)\rangle_{L^2(P;\mathbb{R})}\right)_{i=1}^{n_2}\right\|_2 \\
&\leq \frac{\sigma_A}{\sqrt{n_1}}\left\|\left(\|(\partial_f R)_i\|_{L^2(P;\mathbb{R})}\|\sigma((z_1)_j)\|_{L^2(P;\mathbb{R})}\right)_{i=1}^{n_2}\right\|_2 \\
&= \frac{\sigma_A}{\sqrt{n_1}}\left\|\left(\|(\partial_f R)_i\|_{L^2(P;\mathbb{R})}\right)_{i=1}^{n_2}\right\|_2 \|\sigma((z_1)_j)\|_{L^2(P;\mathbb{R})} \\
&= \frac{\sigma_A}{\sqrt{n_1}}\|\partial_f R\|_{L^2(P;\mathbb{R}^{n_2})}\|\sigma((z_1)_j)\|_{L^2(P;\mathbb{R})}.
\end{aligned}$$

Next, let $\theta_p^{(1)}$ be a pth parameter among A_1 and b_1. Then,

$$\begin{aligned}
\frac{\partial \theta_p^{(1)}}{\partial t}(t) &= -\left\langle \partial_f R\big|_{f_{\theta(t)}}, \frac{\partial f_\theta}{\partial \theta_p^{(1)}} \right\rangle_{L^2(P;\mathbb{R}^{n_2})} \\
&= -\frac{\sigma_A}{\sqrt{n_1}}\left\langle \partial_f R\big|_{f_{\theta(t)}}, A_2\,\mathrm{diag}(\sigma'(z_1))\frac{\partial z_1}{\partial \theta_p^{(1)}}(\cdot;t) \right\rangle_{L^2(P;\mathbb{R}^{n_2})} \\
&= -\frac{\sigma_A}{\sqrt{n_1}}\left\langle \partial_f R, A_2\,\mathrm{diag}(\sigma'(z_1))\frac{\partial z_1}{\partial \theta_p^{(1)}} \right\rangle_{L^2(P;\mathbb{R}^{n_2})}
\end{aligned}$$

Using calculations analogous to what we did to obtain the kernel gradient flow, we get

$$\frac{d}{dt}z_1(t) = -\frac{\sigma_A}{\sqrt{n_1}}L_{\Theta^{(1),\infty}\otimes I_{n_1}}\left[\sigma'(z_1)A_2^\intercal \partial_f R\right]$$

and

$$\frac{d}{dt}(z_1)_j(t) = -\frac{\sigma_A}{\sqrt{n_1}}L_{\Theta^{(1),\infty}}[\sigma'((z_1)_j)(A_2)_{:,j}^\intercal \partial_f R].$$

Let $\kappa = \sup_{x\in\mathbb{R}} |\sigma'(x)|$. Then,

$$\begin{aligned}
\frac{d}{dt}\|(z_1)_j(t) - (z_1)_j(0)\|_{L^2(P;\mathbb{R})} &\leq \frac{\sigma_A}{\sqrt{n_1}} \left\|L_{\Theta^{(1),\infty}}[\sigma'((z_1)_j)(A_2)_{:,j}^\intercal \partial_f R]\right\|_{L^2(P;\mathbb{R})} \\
&\leq \frac{\sigma_A}{\sqrt{n_1}} \|L_{\Theta^{(1),\infty}}\|_{\text{op}} \left\|\sigma'((z_1)_j)(A_2)_{:,j}^\intercal \partial_f R\right\|_{L^2(P;\mathbb{R})} \\
&\leq \frac{\sigma_A}{\sqrt{n_1}} \kappa \|L_{\Theta^{(1),\infty}}\|_{\text{op}} \left\|(A_2)_{:,j}^\intercal \partial_f R(\cdot)\right\|_{L^2(P;\mathbb{R})} \\
&\leq \frac{\sigma_A}{\sqrt{n_1}} \kappa \|L_{\Theta^{(1),\infty}}\|_{\text{op}} \left\|\|(A_2)_{:,j}\|_2 \|\partial_f R(\cdot)\|_2\right\|_{L^2(P;\mathbb{R})} \\
&= \frac{\sigma_A}{\sqrt{n_1}} \kappa \|L_{\Theta^{(1),\infty}}\|_{\text{op}} \|(A_2)_{:,j}\|_2 \left\|\|\partial_f R(\cdot)\|_2\right\|_{L^2(P;\mathbb{R})} \\
&= \frac{\sigma_A}{\sqrt{n_1}} \kappa \|L_{\Theta^{(1),\infty}}\|_{\text{op}} \|(A_2)_{:,j}\|_2 \left\|\partial_f R(\cdot)\right\|_{L^2(P;\mathbb{R}^{n_2})}.
\end{aligned}$$

Again, let $\kappa = \sup_{x\in\mathbb{R}} |\sigma'(x)|$ and define

$$\mathcal{E}(t) = \|\sigma((z_1)_j(0))\|_{L^2(P;\mathbb{R})} + \kappa\|(z_1)_j(t) - (z_1)_j(0)\|_{L^2(P;\mathbb{R})} + \|(A_2)_{:,j}(0)\|_2 + \|(A_2)_{:,j}(t) + (A_2)_{:,j}(0)\|_2.$$

Then

$$\mathcal{E}(t) \geq \|\sigma((z_1)_j(t))\|_{L^2(P;\mathbb{R})} + \|(A_2)_{:,j}(t)\|_2$$

and

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}(t) &\leq \frac{\sigma_A}{\sqrt{n_1}} \left(\|\sigma((z_1)_j)\|_{L^2(P;\mathbb{R})} + \kappa^2 \|(A_2)_{:,j}\|_2 \|L_{\Theta^{(1),\infty}}\|_{\text{op}}\right) \|\partial_f R\|_{L^2(P;\mathbb{R}^{n_2})} \\
&\leq \frac{\sigma_A}{\sqrt{n_1}} \max\{\kappa^2 \|L_{\Theta^{(1),\infty}}\|_{\text{op}}, 1\} \|\partial_f R\|_{L^2(P;\mathbb{R}^{n_2})} \mathcal{E}(t).
\end{aligned}$$

Then we apply Grönwall's lemma to get

$$\mathcal{E}(t) \leq \mathcal{E}(0) \exp\left(\frac{\sigma_A}{\sqrt{n_1}} \max\{\kappa^2 \|L_{\Theta^{(1),\infty}}\|_{\text{op}}, 1\} \int_0^t \|\partial_f R\|_{L^2(P;\mathbb{R}^{n_2})}\, dt.\right)$$

Therefore,

$$\mathcal{E}(t) - \mathcal{E}(0) \leq \mathcal{O}_p(1/\sqrt{n_1}),$$

so

$$\|(z_1)_j(t) - (z_1)_j(0)\|_{L^2(P;\mathbb{R})} \leq \mathcal{O}_p(1/\sqrt{n_1})$$

and

$$\|(A_2)_{:,j}(t) - (A_2)_{:,j}(0)\|_2 \leq \mathcal{O}_p(1/\sqrt{n_1})$$

for all $j = 1, \ldots, n_1$.

Thus, we have

$$
\begin{aligned}
&\delta_{k'k} \left\| \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \sigma((z_1)_j(x))\sigma((z_1)_j(x')) - \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \sigma((z_1)_j(x))\sigma((z_1)_j(x')) \Big|_{t=0} \right\| \\
&\leq \delta_{k'k} \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \left\| \sigma((z_1)_j(x))\sigma((z_1)_j(x')) - \sigma((z_1)_j(x))\sigma((z_1)_j(x')) \Big|_{t=0} \right\| \\
&\leq \delta_{k'k} \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \left\| \sigma((z_1)_j(x))\sigma((z_1)_j(x')) - \sigma((z_1)_j(x))\sigma((z_1)_j(x')) \Big|_{t=0} \right\| \\
&\leq \delta_{k'k} \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \left(\left\| (\sigma((z_1)_j(x)) - \sigma((z_1)_j(x)) \Big|_{t=0}) \sigma((z_1)_j(x')) \right\| \right. \\
&\left. + \left\| \sigma((z_1)_j(x))(\sigma((z_1)_j(x')) - \sigma((z_1)_j(x'))) \Big|_{t=0} \right\| \right) \\
&= \delta_{k'k} \frac{\sigma_A^2}{n_1} \sum_{j=1}^{n_1} \mathcal{O}_p(1/\sqrt{n_1}) \to 0
\end{aligned}
$$

□

5.5 Quadratic case

Again, consider the empirical risk minimization with data $X_1, \ldots, X_N \in \mathcal{X}$, empirical distribution

$$P_{\text{emp}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i},$$

and quadratic loss

$$
\begin{aligned}
R_{\text{emp}}[f] &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \| f(X_i) - f^\star(X_i) \|^2 \\
&= \frac{1}{2} \mathbb{E}_{X \sim P_{\text{emp}}} \left[\| f(x) - f^\star(x) \|^2 \right] \\
&= \frac{1}{2} \langle f - f^\star, f - f^\star \rangle_{L^2(P_{\text{emp}}; \mathbb{R}^{n_L})}.
\end{aligned}
$$

Then

$$\partial_f R_{\text{emp}} = f - f^\star \in L^2(P_{\text{emp}}; \mathbb{R}^{n_L}).$$

Consider the limiting kernel gradient flow with $\Theta = \Theta^{(L),\infty} \times I_{n_L}$:

$$\frac{d}{dt} f_t = -L_t[f_\theta - f^\star],$$

the dynamics is described by

$$f_t = f^\star + e^{-tL_\Theta}[f_t - f^\star].$$

(See homework 5.) In particular, if Θ is strictly positive definite, then

$$f_t(X_i) \to f^\star(X_i), \qquad \text{for } i = 1, \dots, N.$$

Finally, we check the stochastic boundedness assumption. For finite width, we have

$$\frac{d}{dt} f_{\theta(t)} = -L_{\Theta_t^{(L)}}[f_{\theta(t)} - f^\star],$$

with the time-dependent kernel $\Theta_t^{(L)}$, and we still have

$$\frac{d}{dt} R_{\text{emp}}[f_{\theta(t)}] \leq 0.$$

Since

$$\|\partial_f R_{\text{emp}}|_{f_{\theta(t)}}\| = \sqrt{2R_{\text{emp}}[f_{\theta(t)}]} \leq \sqrt{2R_{\text{emp}}[f_{\theta(0)}]},$$

the value

$$\int_0^T \|\partial_f R_{\text{emp}}\| \, dt \leq T\sqrt{2R_{\text{emp}}[f_{\theta(0)}]}$$

is stochastically bounded as $n_1, \dots, n_{L-1} \to \infty$, for fixed $T > 0$, if $R_{\text{emp}}[f_{\theta(0)}]$ is stochastically bounded as $n_1, \dots, n_{L-1} \to \infty$. This is the case, since

$$R_{\text{emp}}[f_{\theta(0)}] \to \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \|f(X_i) - f^\star(X_i)\|^2$$

in distribution with $f \in \mathcal{GP}(0, \Sigma^{(L)})$, and convergence in distribution implies stochastic boundedness.

Chapter 6

Wasserstein distance

Let $\Theta \subseteq \mathbb{R}^p$ and $\Phi \subseteq \mathbb{R}^p$. Write $\mathcal{P}(\Theta)$ and $\mathcal{P}(\Phi)$ for the spaces of probability measures on Θ and Φ, respectively. Write $\mathcal{P}^p(\Theta) \subseteq \mathcal{P}(\Theta)$ and $\mathcal{P}^p(\Phi) \subseteq \mathcal{P}(\Phi)$ for probability measures with finite pth moment. For $p \in [1, \infty)$, the pth Wasserstein distance

$$W_p \colon \mathcal{P}^p(\Theta) \times \mathcal{P}^p(\Phi) \to \mathbb{R}$$

is defined as

$$W_p(\mu, \nu) = \left(\mathbb{E}[\|\theta - \phi\|^p]\right)^{1/p}.$$

We are primarily interested in the W_1, as it is used in WGANs and other deep learning setups due to its Kantorovich–Rubinstein duality variational formulation, and W_2 as it is used in the mean-field theory.

6.1 Optimal transport formulations

We start with the general setup in which Θ and Φ be nonempty metric spaces and add further assumptions on Θ and Φ when necessary. In this section, we describe the mathematical formulations for "transporting" $\mu \in \mathcal{P}(\Theta)$ to $\nu \in \mathcal{P}(\Phi)$.

6.1.1 Monge formulation

Let $T \colon \Theta \to \Phi$ be measurable. We say T is a transport map from μ to ν if $\nu = T_{\#}\mu$. One can picture μ as a pile of sand and grains of sand at $\theta \in \Theta$ are transported to $T(\theta) \in \Phi$. The pushforward measure $\nu = T_{\#}\mu$ represents the resulting pile of sand.

Pushforward measure. We briefly review the notion of the pushforward measure. The pushforward of μ with respect to T is denotes as

$$T_{\#}\mu = \mu \circ T^{-1}$$

and is defined by

$$(T_{\#}\mu)(B) = (\mu \circ T^{-1})(B) = \mu(T^{-1}(B))$$

for all ν-measurable B. The integral with respect to $\nu = T_{\#}\mu$ can be understood via the change-of-variables formula

$$\int_{\Theta} f(T(\theta))\, d\mu(\theta) = \int_{\Phi} f(\phi)\, d\nu(\phi).$$

To put it differently,

$$\mathbb{E}_{\theta\sim\mu}[f(T(\theta))] = \mathbb{E}_{\phi\sim\nu}[f(\phi)],$$

i.e., if $\theta \sim \mu$ and $\phi = T(\theta)$, then $\phi \sim \nu$.

Monge's optimal transport problem is

$$\begin{array}{ll} \underset{T\colon \Theta\to\Phi}{\text{minimize}} & \displaystyle\int_{\Theta} c(\theta, T(\theta))\, d\mu(\theta), \\ \text{subject to} & \nu = T_{\#}\mu, \end{array}$$

where $c(\theta, \phi)$ is cost of transporting mass from θ to ϕ. The particular setups of interest to us are $\Theta = \Phi \subseteq \mathbb{R}^d$ and $c(\theta, \phi) = \|\theta - \phi\|$ or $c(\theta, \phi) = \|\theta - \phi\|^2$. One can picture incurring a cost of

$$c(\theta, T(\theta)) \times (\text{mass of grain})$$

for transporting the grain of sand from θ to $T(\theta)$, and the integral corresponds to the sum of the incurred costs for all grains of sand. The constraint $\nu = T_{\#}\mu$ corresponds to the requirement that the initial profile of sand μ must become ν after the transport is completed.

This formulation is, while easily interpretable, difficult to work with as the optimization problem is non-convex. (Convexity is useful not just for obtaining a computationally efficient algorithm; convexity also provides many theoretically favorable properties.) Furthermore, an optimal transport map T will in general not exist. An easy counterexample is $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$.

Example: Book shifting. Consider two discrete measure $\mu = \frac{1}{N}\sum_{i=1}^{N}\delta_i$ and $\nu = \frac{1}{N}\sum_{i=1}^{N}\delta_{i+1}$ with $\Theta = \Phi = \mathbb{R}$ and $c(\theta,\phi) = \|\theta - \phi\|$. Here are two optimal transport maps:

1. $T(\theta) = \theta$ for $\theta = 2, \cdots, N$ and $T(1) = N+1$
2. $T(\theta) = \theta + 1$ for $\theta = 1, \cdots, N$.

Thus, the optimal transport map need not be unique. (While optimality should be intuitively clear, you will rigorously establish it via duality in the homework.) Also, the value of T outside of $\mathrm{supp}(\mu)$ is irrelevant.

Here is a suboptimal transport map:

$$T(\theta) = \begin{cases} \theta + 2 & \text{for } \theta = 1, \cdots, N-1 \\ 2 & \text{for } \theta = N. \end{cases}$$

So not all (feasible) transport maps are optimal.

6.1.2 Kantorovich formulation

We now consider Kantorovich's relaxed formulation. Consider a measure

$$\pi \in \mathcal{P}(\Theta \times \Phi)$$

Let $P_\Theta \colon \Theta \times \Phi \to \Theta$ be $P_\Theta(\theta,\phi) = \theta$ and $P_\Phi \colon \Theta \times \Phi \to \Phi$ be $P_\Phi(\theta,\phi) = \phi$ be the projection operators. Then $P_{\Theta\#}\pi$ and $P_{\Phi\#}\pi$ are the marginals in the sense that

$$\int_{\Theta\times\Phi} f(\theta)\, d\pi(\theta,\phi) = \int_\Theta f(\theta)\, d(P_{\Theta\#}\pi)(\theta)$$

and

$$\int_{\Theta\times\Phi} f(\phi)\, d\pi(\theta,\phi) = \int_\Phi f(\phi)\, d(P_{\Phi\#}\pi)(\phi).$$

Kantorovich's formulation of the optimal transport problem is

$$\begin{array}{ll} \underset{\pi\in\mathcal{P}(\Theta\times\Phi)}{\text{minimize}} & \displaystyle\int_{\Theta\times\Phi} c(\theta,\phi)\, d\pi(\theta,\phi), \\ \text{subject to} & P_{\Theta\#}\pi = \mu \\ & P_{\Phi\#}\pi = \nu. \end{array}$$

In a probabilistic formulation, one can equivalently write

$$\begin{array}{ll} \underset{\pi\in\mathcal{P}(\Theta\times\Phi)}{\text{minimize}} & \mathbb{E}_{(\theta,\phi)\sim\pi}[c(\theta,\phi)], \\ \text{subject to} & \theta \sim \mu \\ & \phi \sim \nu. \end{array}$$

Since $\pi = \mu \otimes \nu$ is a feasible point of the optimization problem, the optimal value is less than $+\infty$, provided that $\int_{\Theta\times\Phi} c(\theta,\phi)\, d\mu(\theta)d\nu(\phi) < \infty$. If $c(\cdot,\cdot) \geq 0$, then the optimal value is greater than $-\infty$.

Theorem 42 (Disintegration theorem)**.** *Let $\pi \in \mathcal{P}(\Theta \times \Phi)$ and $\mu = P_{\Theta\#}\pi$. Then, there exists a family of conditional probability measures $\{\tilde{\nu}_\theta\}_{\theta\in\Theta} \subset \mathcal{P}(\Phi)$ (defined for μ-almost all θ) such that*

$$d\pi(\theta,\phi) = d\tilde{\nu}_\theta(\phi)d\mu(\theta),$$

i.e.,

$$\int_{\Theta\times\Phi} f(\theta,\phi)\, d\pi(\theta,\phi) = \int_\Theta \int_\Phi f(\theta,\phi)\, d\tilde{\nu}_\theta(\phi)d\mu(\theta), \qquad \forall \text{ measurable } f.$$

The disintegration theorem (which actually has a more general form) ensures that conditional probability measures are well defined. We can interpret the disintegration

$$\int_{\Theta\times\Phi} c(\theta,\phi)\, d\pi(\theta,\phi) = \int_\Theta \underbrace{\int_\Phi c(\theta,\phi)\, d\tilde{\nu}_\theta(\phi)}_{=\substack{\text{unit cost of transporting}\\ \text{sand at } \theta \text{ to } d\tilde{\nu}_\theta(\phi)}} d\mu(\theta),$$

as distributing the sand at θ (we have $d\mu(\theta)$ of sand at θ) to Φ with proportion $d\tilde{\nu}_\theta(\phi)$. Therefore, if $T\colon \Theta \to \Phi$ is feasible for the Monge formulation, then $\delta_{T(\theta)}(d\phi)\mu(d\theta)$ is feasible for for Kantorovich formulation. Thus the Kantorovich formulation has a smaller optimal value than the Monge formulation.

As an aside, if $\pi = \mu \otimes \nu$, then the disintegration is

$$d\pi(\theta,\phi) = d\nu(\phi)d\mu(\theta),$$

and the transport by π takes any infinitesimal grain of sand, regardless of which position θ it came from, split and distributes it according to profile $\nu(\phi)$. This is likely inefficient, as you would probably want to transport the grain of sand differently depending on which initial position θ it came from.

While Monge's formulation is arguably more interpretable, Kantorovich's formulation has better theoretical properties: Kantorovich's formulation often has a solution when Monge's doesn't and Kantorovich's formulation leads to a nice duality framework. In any case, the two formulations yield the same optimal transport cost under the following generic condition.

Theorem 43. *If μ has no atoms, i.e., $\mu(\{\theta\}) = 0$ for all $\theta \in \Theta$, then the optimal values of the Monge and Kantorovich's formulation are the same. (However, it is possible for the Kantorovich formulation to have a solution while the Monge formulation does not.)*

6.1.3 Wasserstein distance

Let $\Theta \subseteq \mathbb{R}^d$ and $\Phi \subseteq \mathbb{R}^d$. For $p \in [1, \infty)$, define

$$\mathcal{P}^p(\Theta) = \{\mu \in \mathcal{P}(\Theta) \,|\, \mathbb{E}_{\theta\sim\mu}[\|\theta\|^p] < \infty\}$$
$$\mathcal{P}^p(\Phi) = \{\nu \in \mathcal{P}(\Phi) \,|\, \mathbb{E}_{\phi\sim\nu}[\|\phi\|^p] < \infty\}.$$

For $p \in [1, \infty)$, define the pth Wasserstein distance $W_p \colon \mathcal{P}^p(\Theta) \times \mathcal{P}^p(\Phi) \to \mathbb{R}$ as

$$(W_p(\mu,\nu))^p = \left(\begin{array}{ll} \underset{\pi\in\mathcal{P}(\Theta\times\Phi)}{\text{minimize}} & \displaystyle\int_{\Theta\times\Phi} \|\theta-\phi\|^p \, d\pi(\theta,\phi), \\ \text{subject to} & P_{\Theta\#}\pi = \mu \\ & P_{\Phi\#}\pi = \nu. \end{array} \right).$$

Theorem 44. *Let $\Theta = \Phi \subseteq \mathbb{R}^d$ and $p \in [1, \infty)$. Then W_p is a metric on $\mathcal{P}^p(\Theta)$.*

Proof. First, we consider the three explicit axioms of metrics. That $W_p(\mu, \nu) = 0$ if and only if $\mu = \nu$ is clear from the probabilistic formulation. That $W_p(\mu, \nu) = W_p(\nu, \mu)$ is also clear. The triangle inequality

$$W_p(\lambda, \nu) \leq W_p(\lambda, \mu) + W_p(\mu, \nu), \qquad \forall\, \lambda, \mu, \nu \in \mathcal{P}^p(\Theta)$$

follows from an argument based on the disintegration theorem. We defer the argument to a homework assignment.

Finally, we verify the fourth implicit axiom of a metric, that $W_p(\mu, \nu) < \infty$ for any $\mu, \nu \in \mathcal{P}^p(\Theta)$. This follows from

$$\begin{aligned} \int_{\Theta\times\Phi} \|\theta-\phi\|^p \, d\pi(\theta,\phi) &\leq \int_{\Theta\times\Phi} 2^{p-1}\|\theta\|^p + 2^{p-1}\|\phi\|^p \, d\pi(\theta,\phi) \\ &= 2^{p-1}\int_\Theta \|\theta\|^p \, d\mu(\theta) + 2^{p-1}\int_\Phi \|\phi\|^p \, d\nu(\phi) \\ &< \infty. \end{aligned}$$

□

In intuitive terms, the triangle inequality is a very natural conclusion. When transporting from λ to ν, you always have the option of transporting from λ to μ and then from μ to ν. Going through the intermediate step μ should increase the cost.

6.2 Duality

As the Wasserstein distances are defined through (infinite-dimensional) constrained convex optimization problems, they have a rich duality theory. The the dual of W_1, specifically referred to as the Kantorovich–Rubinstein dual, is used commonly in modern deep learning, most notably in the WGAN. The dual of W_2 leads to Brenier's theorem, which is used in the characterization of Wasserstein gradient flows and the formulation of the mean-field limit of wide 2-layer neural networks.

Let Ω and Φ be nonempty metric spaces. Let $c\colon \Omega\times\Phi \to \mathbb{R}_+$. For $\mu \in \mathcal{P}(\Omega)$ and $\nu \in \mathcal{P}(\Phi)$, define

$$W(\mu,\nu) = \left(\begin{array}{ll} \underset{\pi\in\mathcal{P}(\Theta\times\Phi)}{\text{minimize}} & \displaystyle\int_{\Theta\times\Phi} c(\theta,\phi)\, d\pi(\theta,\phi), \\ \text{subject to} & P_{\Theta\#}\pi = \mu \\ & P_{\Phi\#}\pi = \nu. \end{array} \right).$$

Define $W(\mu,\nu) = \infty$ if there is no $\pi \in \mathcal{P}(\Theta\times\Phi)$ that makes the integral finite. Given a measurable function f and a measure μ, write $\langle f, \mu\rangle = \langle \mu, f\rangle = \int f d\mu$ to denote the corresponding integral.

Let

$$\mathbf{L}(\pi,\varphi,\psi) = \langle c, \pi\rangle + \langle \mu - P_{\Theta\#}\pi, \varphi\rangle + \langle \nu - P_{\Phi\#}\pi, \psi\rangle.$$

Note that $\mathbf{L}$ is convex in π and concave in (φ,ψ). (In fact, linear in π and linear in (φ,ψ).) Define the primal problem

$$\overline{W(\mu,\nu)} := \inf_{\pi\in\mathcal{M}_+(\Theta\times\Phi)} \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \mathbf{L}(\pi,\varphi,\psi)$$

and the dual problem

$$\underline{W(\mu,\nu)} := \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \inf_{\pi\in\mathcal{M}_+(\Theta\times\Phi)} \mathbf{L}(\pi,\varphi,\psi) \le W(\mu,\nu).$$

Using the general weak duality principle, we get

$$\sup\inf \boxed{} \le \inf\sup \boxed{}.$$

We let $\varphi \in \mathcal{C}_0(\Theta)$ because $(\mathcal{C}_0(\Theta), \|\cdot\|_\infty)$ is the pre-dual space of $(\mathcal{M}(\Theta), \|\cdot\|_{\mathrm{TV}})$. The same reason holds for $\psi \in \mathcal{C}_0(\Phi)$. This careful pairing of (topological) dual spaces is necessary for applying the Fenchel–Rockafellar duality theorem and obtaining strong duality (although we omit this discussion).

Primal problem. The primal problem recovers the Kantorovich formulation

$$\begin{aligned}\overline{W(\mu,\nu)} &= \inf_{\pi\in\mathcal{M}_+(\Theta\times\Phi)} \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \langle c,\pi\rangle + \langle \mu - P_{\Theta\#}\pi,\varphi\rangle + \langle \nu - P_{\Phi\#}\pi,\psi\rangle \\ &= \inf_{\pi\in\mathcal{M}_+(\Theta\times\Phi)} \begin{cases} \langle c,\pi\rangle & \text{if } \mu - P_{\Theta\#}\pi = 0,\ \nu - P_{\Phi\#}\pi = 0 \\ \infty & \text{otherwise.}\end{cases} \\ &= W(\mu,\nu).\end{aligned}$$

Here, we use the fact that $\mu = P_{\Theta\#}\pi$ implies that π has total mass 1, i.e., $\pi \in \mathcal{M}_+(\Theta\times\Phi)$ and $P_{\Theta\#}\pi \in \mathcal{P}(\Theta)$ implies $\pi \in \mathcal{P}(\Theta\times\Phi)$.

Dual problem. Consider

$$\begin{aligned}\mathbf{L}(\pi,\varphi,\psi) &= \langle c,\pi\rangle - \langle \pi, P^\dagger_{\Theta\#}\varphi\rangle - \langle \pi, P^\dagger_{\Phi\#}\psi\rangle + \langle \mu,\varphi\rangle + \langle \nu,\psi\rangle \\ &= \langle c - P^\dagger_{\Theta\#}\varphi - P^\dagger_{\Phi\#}\psi,\pi\rangle + \langle \mu,\varphi\rangle + \langle \nu,\psi\rangle\end{aligned}$$

with $\langle P_{\Theta\#}\pi,\varphi\rangle = \int_\Theta \varphi d(P_{\Theta\#}\pi)(\theta) = \int_{\Theta\times\Phi}\varphi(\theta)d\pi(\theta,\phi) = \langle \pi, P^\dagger_{\Theta\#}\varphi\rangle$. So we can view the adjoint operator $P^\dagger_{\Theta\#}\colon \mathcal{C}_0(\Theta)\to\mathcal{C}_0(\Theta\times\Phi)$ as the pedantic operation mapping $\varphi(\theta)$ and $\varphi(\theta,\phi) = \varphi(\theta)$. In less abstract notation, we have

$$\mathbf{L}(\pi,\varphi,\psi) = \int_{\Theta\times\Phi} c(\theta,\phi) - \varphi(\theta) - \psi(\phi)\ d\pi(\theta,\phi) + \int_\Theta \varphi(\theta)\ d\mu(\theta) + \int_\Phi \psi(\phi)\ d\mu(\phi).$$

We now characterize the dual problem

$$\begin{aligned}\underline{W(\mu,\nu)} &= \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \inf_{\pi\in\mathcal{M}_+(\Theta\times\Phi)} \langle c - P^\dagger_{\Theta\#}\varphi - P^\dagger_{\Phi\#}\psi,\pi\rangle + \langle\mu,\varphi\rangle + \langle\nu,\psi\rangle \\ &= \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \begin{cases} \langle\mu,\varphi\rangle + \langle\nu,\psi\rangle & \text{if } c - P^\dagger_{\Theta\#}\varphi - P^\dagger_{\Phi\#}\psi \geq 0 \\ -\infty & \text{otherwise.}\end{cases}\end{aligned}$$

To clarify,

$$P^\dagger_{\Theta\#}\varphi + P^\dagger_{\Phi\#}\psi \leq c$$

means

$$\varphi(\theta) + \psi(\phi) \leq c(\theta,\phi), \qquad \forall\,\theta\in\Theta,\ \phi\in\Phi.$$

Finally, we write

$$\underline{W(\mu,\nu)} = \left(\begin{array}{ll} \underset{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)}{\text{maximize}} & \displaystyle\int_\Theta \varphi(\theta)\ d\mu(\theta) + \int_\Phi \psi(\phi)\ d\nu(\phi), \\ \text{subject to} & \varphi(\theta) + \psi(\phi) \leq c(\theta,\phi), \qquad \forall\,\theta\in\Theta,\ \phi\in\Phi. \end{array}\right)$$

If μ and ν are compactly supported, then equivalently have

$$\underline{W(\mu,\nu)} = \left(\begin{array}{ll} \underset{\varphi\in\mathcal{C}(\Theta),\psi\in\mathcal{C}(\Phi)}{\text{maximize}} & \displaystyle\int_\Theta \varphi(\theta)\ d\mu(\theta) + \int_\Phi \psi(\phi)\ d\nu(\phi), \\ \text{subject to} & \varphi(\theta) + \psi(\phi) \leq c(\theta,\phi), \qquad \forall\,\theta\in\Theta,\ \phi\in\Phi. \end{array}\right)$$

Strong duality and complementary slackness. Through convex duality theory (specifically, using the Fenchel–Rockafellar duality theorem) one can establish strong duality

$$\overline{W(\mu,\nu)} = \underline{W(\mu,\nu)}.$$

In the interest of time, we skip this proof, but we will accept the result and write

$$W(\mu,\nu) = \overline{W(\mu,\nu)} = \underline{W(\mu,\nu)}.$$

An additional useful consequence of strong duality is complementary slackness. If $\pi^\star$ solves the primal problem and $(\varphi^\star, \psi^\star)$ solves the dual problem,

$$\begin{aligned}
\overline{W(\mu,\nu)} &= \inf_{\pi\in\mathcal{M}_+(\Theta\times\Phi)} \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \langle c - P^\dagger_{\Theta\#}\varphi - P^\dagger_{\Phi\#}\psi, \pi\rangle + \langle\mu,\varphi\rangle + \langle\nu,\psi\rangle \\
&= \sup_{\varphi\in\mathcal{C}_0(\Theta),\psi\in\mathcal{C}_0(\Phi)} \langle c - P^\dagger_{\Theta\#}\varphi - P^\dagger_{\Phi\#}\psi, \pi^\star\rangle + \langle\mu,\varphi\rangle + \langle\nu,\psi\rangle \\
&\geq \langle c - P^\dagger_{\Theta\#}\varphi^\star - P^\dagger_{\Phi\#}\psi^\star, \pi^\star\rangle + \langle\mu,\varphi^\star\rangle + \langle\nu,\psi^\star\rangle \\
&\geq \langle c - P^\dagger_{\Theta\#}\varphi^\star - P^\dagger_{\Phi\#}\psi^\star, \pi^\star\rangle + \underline{W(\mu,\nu)}
\end{aligned}$$

Since $W(\mu,\nu) = \overline{W(\mu,\nu)} = \underline{W(\mu,\nu)}$ by strong duality, we have

$$0 \geq \langle \underbrace{c - P^\dagger_{\Theta\#}\varphi^\star - P^\dagger_{\Phi\#}\psi^\star}_{\geq 0}, \underbrace{\pi^\star}_{\geq 0}\rangle,$$

and we conclude complementary slackness:

$$\varphi^\star(\theta) + \psi^\star(\phi) = c(\theta,\phi), \qquad \forall\,(\theta,\phi) \in \mathrm{supp}(\pi^\star).$$

6.2.1 Kantorovich–Rubinstein duality

Assume for simplicity that μ and ν have compact support. Assume $\Theta = \Phi \subseteq \mathbb{R}^d$. We shall obtain the Kantorovich–Rubinstein dual by simplifying the dual problem for W_1:

$$W_1(\mu,\nu) = \left(\begin{array}{ll} \underset{\varphi\in\mathcal{C}(\Theta),\psi\in\mathcal{C}(\Phi)}{\text{maximize}} & \displaystyle\int_\Theta \varphi(\theta)\,d\mu(\theta) + \int_\Phi \psi(\phi)\,d\nu(\phi), \\ \text{subject to} & \varphi(\theta) + \psi(\phi) \leq \|\theta - \phi\|, \qquad \forall\,\theta\in\Theta,\ \phi\in\Phi. \end{array} \right)$$

Let (φ,ψ) be feasible. This implies

$$\psi(\phi) \leq \inf_{\theta\in\Theta}\{\|\theta-\phi\| - \varphi(\theta)\}.$$

Define

$$\varphi^c(\phi) = \inf_{\theta\in\Theta}\{\|\theta-\phi\| - \varphi(\theta)\}.$$

Then, (φ, φ^c) is also feasible, i.e., $\varphi(\theta)+\varphi^c(\phi) \le \|\theta-\phi\|$, and φ^c is the largest feasible ψ given φ. Certainly, $\psi \le \varphi^c$, and replacing (φ,ψ) with (φ,φ^c) can only improve the objective, since ν is a nonnegative measure. Define

$$\varphi^{cc}(\theta) = \inf_{\phi\in\Phi}\{\|\theta-\phi\| - \varphi^c(\phi)\}.$$

Then (φ^{cc},φ^c) is also a feasible point and φ^{cc} is the largest feasible φ given φ^c. Replacing (φ,φ^c) with (φ^{cc},φ^c) can only improve the objective.[1] We can now restrict the optimization problem to

$$\Gamma = \{(\varphi^{cc},\varphi^c)\,|\,\varphi\in\mathcal{C}(\Theta), \varphi^c > -\infty\}$$

(here, $\varphi^c > -\infty$ meaans $\varphi^c(\phi)\neq-\infty$ for all $\phi\in Phi$) and write

$$W_1(\mu,\nu) = \left(\underset{(\varphi^{cc},\varphi^c)\in\Gamma}{\text{maximize}} \int_\Theta \varphi^{cc}(\theta)\,d\mu(\theta) + \int_\Phi \varphi^c(\phi)\,d\nu(\phi)\right).$$

Now we further characterize Γ. Note that $(\varphi^{cc},\varphi^c)\in\Gamma$ are 1-Lipschitz continuous, since

$$\varphi^c(\phi) = \inf_{\theta\in\Theta}\{\|\theta-\phi\| - \varphi(\theta)\}$$

is an infimum of 1-Lipschitz continuous functions, and ditto for φ^{cc}. Conversely, if φ is 1-Lipschitz, then

$$\varphi^c = -\varphi, \qquad \varphi^{cc} = \varphi.$$

This follows from

$$-\varphi(\phi) \le \|\theta-\phi\| - \varphi(\theta),$$

implied by 1-Lipschitz continuity, and

$$\inf_\theta\{\|\theta-\phi\| - \varphi(\theta)\} \le -\varphi(\phi)$$

because the infimum is no larger than simply plugging in $\theta=\phi$. Therefore,

$$\Gamma = \{(\varphi,-\varphi)\,|\,\varphi\in\mathcal{L}_1(\Theta)\}.$$

and we conclude

$$W_1(\mu,\nu) = \left(\begin{array}{ll}\underset{\varphi}{\text{maximize}} & \displaystyle\int_\Theta \varphi(\theta)\,d\mu(\theta) - \int_\Phi \varphi(\phi)\,d\nu(\phi) \\ \text{subject to} & \varphi\in\mathcal{L}_1\end{array}\right).$$

[1] The mapping $\varphi\mapsto\varphi^c$ defined by $\varphi^c(\phi) = \inf_{\theta\in\Theta}\{c(\theta,\phi)-\varphi(\theta)\}$ is called the c-transform and it generalizes the conjugacy of convex functions in some sense. One can show that $\varphi^{ccc}=\varphi^c$, so there is no need to further continue this process.

6.2.2 Preliminaries: Convex conjugates

We say a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex if

$$f(\eta x + (1-\eta)y) \le \eta f(x) + (1-\eta) f(y), \qquad \forall\, x, y \in \mathbb{R}^d,\ \eta \in (0,1).$$

In convex analysis, the class of regular (nice) convex functions most commonly considered is called closed, convex, and proper (CCP) functions. To clarify CCP functions are, in general, extended real-valued, i.e., the output can be ∞.

In this course, however, we will not fully define the notion of CCP, as we do not need it. Rather, we will say a function f is *convex and finite* if it is convex and $f(x) < \infty$. Convex and finite functions are CCP, continuous, and are differentiable almost everywhere.

The convex conjugate of $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(u) = \sup_{x \in \mathbb{R}^d} \{\langle u, x\rangle - f(x)\},$$

where $f^*\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. If f is CCP, then f^* is CCP and $f^{**} = f$. However, even if f is convex and finite, f^* is not guaranteed to be finite.

From the definition of the convex conjugate, we immediately have the Fenchel–Young inequality

$$\langle u, x\rangle \le f^*(u) + f(x), \qquad \forall\, x, u \in \mathbb{R}^d.$$

Also, if f is CCP, f is differentiable at x, if

$$\langle u, x\rangle = f^*(u) + f(x),$$

then $u = \nabla f(x)$. (This does not follow from simply differentiating both sides with respect to x.) To see why,

$$\begin{aligned} \langle u, x\rangle - f(x) &= f^*(u) \\ &= \sup_{z \in \mathbb{R}^d} \{\langle u, z\rangle - f(z)\}, \end{aligned}$$

so $u = \nabla f(x)$.

6.2.3 Brenier's theorem: W_2

Assume for simplicity that μ and ν have compact support and that $\Theta = \Phi = \mathbb{R}^d$. Let us obtain Brenier's theorem by simplifying the dual problem for W_2

$$W_2^2(\mu,\nu) = \left(\begin{array}{ll} \underset{\varphi \in \mathcal{C}(\Theta), \psi \in \mathcal{C}(\Phi)}{\text{maximize}} & \displaystyle\int_\Theta \varphi(\theta)\, d\mu(\theta) + \int_\Phi \psi(\phi)\, d\nu(\phi), \\ \text{subject to} & \varphi(\theta) + \psi(\phi) \le \frac{1}{2}\|\theta - \phi\|^2, \qquad \forall\, \theta \in \Theta,\ \phi \in \Phi \end{array} \right)$$

Let (φ, ψ) be feasible. Define

$$\begin{aligned}\varphi^c(\phi) &= \inf_{\theta\in\Theta}\left\{\frac{1}{2}\|\theta-\phi\|^2 - \varphi(\theta)\right\} \\ &= -\sup_{\theta\in\Theta}\left\{\langle\theta,\phi\rangle - \left(\frac{1}{2}\|\theta\|^2 - \varphi(\theta)\right)\right\} + \frac{1}{2}\|\phi\|^2 \\ &= -\left(\frac{1}{2}\|\cdot\|^2 - \varphi\right)^*(\phi) + \frac{1}{2}\|\phi\|^2,\end{aligned}$$

where $*$ denotes the convex conjugate. Likewise, define

$$\begin{aligned}\varphi^{cc}(\theta) &= \inf_{\phi\in\Phi}\left\{\frac{1}{2}\|\theta-\phi\|^2 - \varphi^c(\phi)\right\} \\ &= -\left(\frac{1}{2}\|\cdot\|^2 - \varphi^c\right)^*(\theta) + \frac{1}{2}\|\theta\|^2 \\ &= -\left(\frac{1}{2}\|\cdot\|^2 - \varphi\right)^{**}(\theta) + \frac{1}{2}\|\theta\|^2.\end{aligned}$$

From the same reasoning as before, $(\varphi^{cc}, \varphi^c)$ is feasible, and it achieves an objective value no worse than (φ, ψ). Define

$$\Gamma = \{(\varphi^{cc}, \varphi^c) \,|\, \varphi \in \mathcal{C}(\Theta), \varphi^c > -\infty\}$$

and write

$$W_2^2(\mu,\nu) = \left(\underset{(\varphi^{cc},\varphi^c)\in\Gamma}{\text{maximize}} \int_\Theta \varphi^{cc}(\theta)\, d\mu(\theta) + \int_\Phi \varphi^c(\phi)\, d\nu(\phi)\right).$$

Now we further characterize Γ. Let $\varphi \in \mathcal{C}(\Theta)$, such that $\varphi^c > -\infty$. Then, $\left(\frac{1}{2}\|\cdot\|^2 - \varphi\right)^*$ is convex (since it is a conjugate function) and finite (since $\varphi^c(\phi) > -\infty$). Define $\tau = \left(\frac{1}{2}\|\cdot\|^2 - \varphi\right)^{**}$, which is also convex (since it is a conjugate function) and finite (since $\varphi^{cc} \geq \varphi > -\infty$). Conversely, let τ be convex and finite function such that τ^* is also finite. Define $\varphi = \frac{1}{2}\|\cdot\|^2 - \tau$. Then, $\varphi^c = \frac{1}{2}\|\cdot\|^2 - \tau^*$ and

$$\varphi(\theta) + \varphi^c(\phi) - \frac{1}{2}\|\theta-\phi\|^2 = \langle\theta,\phi\rangle - \tau(\theta) - \tau^*(\phi) \leq 0$$

by Fenchel–Young. Therefore,

$$\Gamma = \left\{\left(\frac{1}{2}\|\cdot\|^2 - \tau, \frac{1}{2}\|\cdot\|^2 - \tau^*\right) \,\middle|\, \tau \text{ convex, } \tau \text{ and } \tau^* \text{ finite}\right\}.$$

and

$$W_2^2(\mu,\nu) = \left(\begin{array}{ll} \underset{\tau\text{: convex}}{\text{maximize}} & \displaystyle\int_\Theta \varphi(\theta)\, d\mu(\theta) + \int_\Phi \psi(\phi)\, d\nu(\phi), \\ \text{subject to} & \varphi = \frac{1}{2}\|\cdot\|^2 - \tau, \quad \tau \text{ finite} \\ & \psi = \frac{1}{2}\|\cdot\|^2 - \tau^*, \quad \tau^* \text{ finite} \end{array} \right).$$

Transport plan. Assume μ is absolutely continuous (with respect to the Lebesgue measure). Then, τ is differentiable μ-almost everywhere. Let π_{opt} and τ_{opt} be optimal primal and dual solutions for W_2. We claim that

$$T = \nabla\tau_{\text{opt}}$$

is an optimal transport plan.

By complementary slackness,

$$\varphi_{\text{opt}}(\theta) + \psi_{\text{opt}}(\phi) = c(\theta,\phi) \quad \Leftrightarrow \quad \langle\theta,\phi\rangle = \tau_{\text{opt}}(\theta) + \tau^*_{\text{opt}}(\phi)$$

π_{opt}-almost everywhere, which implies

$$\nabla\tau_{\text{opt}}(\theta) = \phi$$

π_{opt}-almost everywhere. Therefore,

$$\text{supp}(\pi_{\text{opt}}) \subseteq \{(\theta, \nabla\tau_{\text{opt}}(\theta)) \,|\, \theta \in \Theta\}$$

and the disintegration of π simplifies to

$$d\pi_{\text{opt}}(\theta,\phi) = d\mu(\theta) d\delta_{\nabla\tau_{\text{opt}}(\theta)}(\phi)$$

for the optimal τ_{opt}. Therefore, π_{opt} corresponds to an optimal transport map $T = \nabla\tau_{\text{opt}}$.

Chapter 7

Weak solution of differential equations and Wasserstein gradient flow

7.1 Weak solution to ODE

Consider the ODE

$$\dot{X} = f(t, X), \qquad X(0) = X_0.$$

If $f\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is continuous, then $X(t) \in \mathcal{C}^1([0,\infty); \mathbb{R}^d)$, and hence a solution is well defined. But what is the definition of a solution? Obviously, if you can plug in the solution to the differential equation and it verifies, then it is a solution.

However, what if f is discontinuous? For example, if

$$f(t, X) = \begin{cases} 0 & \text{for } t \leq 1 \\ 1 & \text{otherwise,} \end{cases}$$

then

$$X(t) = \begin{cases} 0 & \text{for } t \leq 1 \\ t - 1 & \text{otherwise.} \end{cases}$$

Right? However, the solution $X(t)$ is non-differentiable at $t = 1$? To generalize the notion of what constitutes a solution, we define X to be a solution if it is an integrable function such that

$$X(t) = X(0) + \int_0^t f(s, X(s))\, ds.$$

We have now completely dropped all requirement that X be differentiable. Now, we never directly differentiate X and we only access f through integration.

Another approach, however, is as follows. For all test functions $\varphi \in \mathcal{C}_c^\infty((0,\infty))$,

$$\int_0^\infty \varphi(t) f(t, X(t))\, dt = -\int_0^\infty \varphi'(t) X(t)\, dt,$$

then X is a solution. (Since φ is compactly supported on $(0,\infty)$, rather than $[0,\infty)$, $\lim_{t\to 0^+} \varphi(t) = 0$.) This formulation via test functions is not yet necessary, but it will be the standard approach for defining weak solutions for PDEs.

7.2 Weak solution to PDE

Consider the following example with the first-order wave equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \qquad u(0,x) = u_0(x) \quad \forall x \in \mathbb{R}.$$

As an example, the initial condition,

$$u(0,x) = e^{-x^2}$$

yields the solution

$$u(t,x) = e^{-(x-t)^2}.$$

We can simply plug in the solution to the PDE and verify that is satisfies the equation. (Solution is unique, but let us not worry about that.) More generally, if

$$u(0,x) = \kappa(x)$$

where $\kappa \in \mathcal{C}^1(\mathbb{R})$. Then

$$u(t,x) = \kappa(x-t)$$

solves the PDE.

However, if

$$u(0,x) = \exp(-|x|),$$

then does

$$u(t,x) = \exp(-|x-t|)$$

solve the PDE? The answer is no. If

$$u(0,x) = \mathbf{1}_{[-1,1]}(x),$$

then does

$$u(t,x) = \mathbf{1}_{[-1,1]}(x-t)$$

solve the PDE? These initial conditions do not yield *strong* solutions to the PDE.

7.2.1 Formal derivation of weak formulation

Assume $u(t,x)$ is a continuously differentiable solution to the PDE, i.e., u is a strong solution. For $\varphi \in \mathcal{C}_c^\infty((0,\infty)\times\mathbb{R})$,

$$\begin{aligned}
0 &= \int_{-\infty}^{\infty}\int_0^{\infty} \varphi(\partial_t u + \partial_x u)\, dtdx \\
&= \int_{-\infty}^{\infty}\int_0^{\infty} \varphi\partial_t u\, dtdx + \int_0^{\infty}\int_{-\infty}^{\infty} \varphi\partial_x u\, dxdt \\
&= -\int_{-\infty}^{\infty}\int_0^{\infty} (\partial_t\varphi) u\, dtdx - \int_0^{\infty}\int_{-\infty}^{\infty} (\partial_x\varphi) u\, dxdt \\
&= -\int_{-\infty}^{\infty}\int_0^{\infty} (\partial_t\varphi + \partial_x\varphi) u\, dtdx
\end{aligned}$$

Therefore, a continuously differentiable function $u(t,x)$ solves the PDE if and only if $u(0,x) = u_0(x)$ and

$$0 = \int_{-\infty}^{\infty}\int_0^{\infty} (\partial_t\varphi + \partial_x\varphi) u\, dtdx, \qquad \varphi \in \mathcal{C}_c^\infty((0,\infty)\times\mathbb{R}).$$

Finally we extend the definition of a solution as follows. We say a (possibly non-differentiable) function $u(t,x)$ is a *weak* solution to the PDE if $u(0,x) = u_0(x)$ and

$$0 = \int_{-\infty}^{\infty}\int_0^{\infty} (\partial_t\varphi + \partial_x\varphi) u\, dtdx, \qquad \varphi \in \mathcal{C}_c^\infty((0,\infty)\times\mathbb{R}).$$

As we have established, a strong solution is a weak solution. However, there are some weak solutions that are not strong solutions.

What we have seen is the general template for defining a weak solution of PDEs. First, assuming the solution is sufficiently smooth, find an equivalent integral characterization of the solution via a smooth, compactly supported test function, often using integration by parts or the divergence theorem. This first step is called the *formal derivation* since we are performing calculations as if the solution is sufficiently smooth (even if it is not) without justifying

whether the formal rules such as integration by parts are actually mathematically valid. Second, we define a weak solution as a function (or a measure) that satisfies the integral form of the equation. Since the integral form was formally derived with mathematically valid rules, if the solution is sufficiently smooth, a sufficiently smooth weak solution is automatically a strong solution.

7.3 Continuity equation

Let $\rho(t,x)$ be a density of particles (e.g. air molecules) at time t and position $x \in \mathbb{R}^d$. Let $\mathbf{v}(t,x;\rho(t,\cdot))$ be the velocity of the particles at position x and with global particle profile $\rho(t,\cdot)$. For now, assume $\rho(t,x)$ is continuously differentiable in t and x and that $\mathbf{v}(t,x;\rho(t,\cdot))$ is continuously differentiable in x. Our formulation informally assumes:

(i) Particles are indistinguishable, i.e., two particles at the same position and time will move under the same velocity. (Say, in a game-theoretic setup, it would be reasonable to consider two agents with different personalities. If so, the two agents would behave differently even under the exact same environment.)

(ii) Particles have long-range interactions, i.e., the $\mathbf{v}$ at (t,x) depends on $\rho(t,\cdot)$, not just $\rho(t,x)$. If particles, say, exert gravitational force, then it would make sense that the dynamics of one particle depends on the global arrangement of other particles.

Note that the global profile can affect the velocity. As an example, if particles exert gravitational attraction, then the force experienced at position x will depend on the global particle profile $\rho(t,\cdot)$. Let $\rho(t,x)\mathbf{v}(t,x)$ be the *flow velocity*. The flow velocity captures the amount of flow. (For points where density ρ is zero, the flow velocity $\rho\mathbf{v}$ becomes zero and the velocity $\mathbf{v}$ becomes irrelevant.)

Many physical systems possess certain conservation laws: mass, charge, and neurons are examples of conserved quantities. In our equations, $\mathbf{u} = \rho\mathbf{v}$ describes the flux of particles. By the conservation law, the change in the number of particle in a volume V is equal to the number of particles escaping through its surface ∂V:

$$\partial_t \int_V \rho \, dV = -\int_{\partial V} \rho(\mathbf{v}\cdot\hat{n}) \, dS = -\int_V \nabla_x \cdot (\rho\mathbf{v}) \, dV,$$

where the first equality follows from the physical modeling and the second equality follows from the divergence theorem. Assume $\partial_t \int_V = \int_V \partial_t$. Since

the volume V is arbitrary, we arrive the continuity equation:

$$\partial_t \rho + \nabla_x \cdot (\rho \mathbf{v}) = 0, \qquad \rho(0, x) = \rho_0(x).$$

The velocity $\mathbf{v} = \mathbf{v}(t, x; \rho(t, \cdot))$ is pre-specified; it is a known function (determined by the physics or SGD algorithm) of t, x, and $\rho(t, \cdot)$. The unknown, defined by the PDE, is $\rho(t, x)$ for $t > 0$.

One use of the continuity equation is to describe the dynamics of (compressible) fluid flow. In this case, $\rho(t, x)$ is the density of particles at (t, x), $\mathbf{v}(t, x; \rho(t, \cdot))$ is the velocity of particles at (t, x), and $\rho\mathbf{v}$ describes the flux of the fluid.

7.3.1 Formal derivation of weak formulation

Assume ρ is continuously differentiable. For $\varphi \in \mathcal{C}_c^\infty((0, \infty) \times \mathbb{R}^d)$,

$$\begin{aligned} 0 &= \int_{\mathbb{R}^d} \int_0^\infty \varphi \, \partial_t \rho \, dt dx + \int_0^\infty \int_{\mathbb{R}^d} \varphi \, \nabla_x \cdot (\rho \mathbf{v}) \, dx dt \\ &= -\int_{\mathbb{R}^d} \int_0^\infty \partial_t \varphi \, \rho \, dt dx - \int_0^\infty \int_{\mathbb{R}^d} \rho((\nabla_x \varphi) \cdot \mathbf{v}) \, dx dt. \end{aligned}$$

(You will show the second equality in a homework assignment.) Thus, we get

$$0 = \int_0^\infty \int_{\mathbb{R}^d} (\partial_t \varphi + \nabla_x \varphi \cdot \mathbf{v}) \, \rho(x, t) \, dx dt, \qquad \forall \, \varphi \in \mathcal{C}_c^\infty((0, \infty) \times \mathbb{R}^d).$$

We now generalize the notion of solution to measures. We say a family of measures $\{\rho_t\}_{t \geq 0} \subset \mathcal{P}(\mathbb{R}^d)$ is a weak solution to the continuity equation

$$\partial_t \rho_t = -\nabla_x \cdot (\mathbf{v}_t \rho_t), \qquad \rho_0 \in \mathcal{P}(\mathbb{R}^d)$$

with $\mathbf{v}_t = \mathbf{v}(t, x; \rho_t)$, if

$$0 = \int_0^\infty \int_{\mathbb{R}^d} (\partial_t \varphi + \nabla_x \varphi \cdot \mathbf{v}_t) \; d\rho_t(x) dt, \qquad \forall \, \varphi \in \mathcal{C}_c^\infty((0, \infty) \times \mathbb{R}^d).$$

In particular, we don't explicitly define the meaning of $\partial_t \rho_t$ or $\nabla_x \cdot (\mathbf{v}_t \rho_t)$. Rather, we view the "differential" equation as a shorthand for the weak formulation.

7.3.2 Properties of the continuity equation

The name "continuity equation" implies that it induces a "continuous flow", and particles or mass are not allowed to "teleport". As an example, let $\eta \in \mathcal{C}_c^\infty(\mathbb{R})$ satisfy $\eta \geq 0$, $\eta(x) = 0$ for $|x| \geq 1/3$, and

$$\int_{-1/3}^{1/3} \eta(x)\, dx = 1.$$

Let

$$\rho(0, x) = \eta(x)$$

Then

$$\rho(t, x) = \eta(x - t), \qquad \text{for } t \in [0, 1],\ x \in \mathbb{R}.$$

is the resulting flow if $\mathbf{v}_t = 1$. However,

$$\rho(t, x) = (1 - t)\eta(x) + t\eta(x - 1), \qquad \text{for } t \in [0, 1],\ x \in \mathbb{R}$$

is not a solution of the continuity equation no matter the choice of $\mathbf{v}$. You will work out the argument as a homework assignment. The argument follows from the use of the divergence theorem.

The following is a more concrete parameterization of the solution of the continuity equation. For each $x \in \mathbb{R}^d$, let $X(\cdot, x)$ be the solution to the ODE

$$\dot{X}(t, x) = \mathbf{v}(t, X(t, x), ; \rho_t) \qquad X(0, x) = x.$$

Then,

$$\rho_t = (X(t, x))_{\#}\rho_0.$$

We can directly verify this. For any $\varphi \in \mathcal{C}_c^\infty((0, \infty) \times \mathbb{R}^d)$,

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^d} (\partial_t \varphi(t, x) + \nabla_x \varphi(t, x) \cdot \mathbf{v}_t(t, x; \rho_t))\ d\rho_t(x)dt \\
&\quad = \int_0^\infty \int_{\mathbb{R}^d} (\partial_t \varphi(t, X(t, x)) + \nabla_x \varphi(t, X(t, x)) \cdot \mathbf{v}_t(t, X(t, x); \rho_t))\ d\rho_0(x)dt \\
&\quad = \int_0^\infty \int_{\mathbb{R}^d} \frac{d}{dt} \varphi(t, X(t, x))\, d\rho_0(x)dt \\
&\quad = \int_{\mathbb{R}^d} \varphi(\infty, X(\infty, x)) - \varphi(0, X(0, x))\, d\rho_0(x) \\
&\quad = 0.
\end{aligned}$$

Under mild assumptions, $X(t, \cdot)\colon \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism. (In ODE theory or differential geometry, the flow of a vector field mapping the initial point x to its position at time t is a diffeomorphism.) In this case, if $\operatorname{supp}(\rho_0) = \mathbb{R}^d$, then $\operatorname{supp}(\rho_t) = \mathbb{R}^d$.

7.4 Wasserstein gradient flow

7.4.1 Metric gradient flow

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be differentiable. As considered in one of the homework problems, gradient flow

$$\dot{x}_t = -\nabla f(x_t)$$

can be defined as the interpolation of

$$\tilde{x}_{t+\varepsilon} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ f(x) + \frac{1}{2\varepsilon}\|x - \tilde{x}_t\|^2 \right\}$$

in the sense that

$$\{\tilde{x}_{\lfloor t/\varepsilon\rfloor\varepsilon}\}_{t\geq 0} \to \{x_t\}_{t\geq 0}$$

as $\varepsilon \to 0$ in the topology of uniform convergence on compacta.

Next, let $f\colon \mathcal{X} \to \mathbb{R}$ where $(\mathcal{X}, d)$ is a metric space. Defining a gradient flow with respect to f initially seems impossible. How can we even define a gradient for f? A most general formulation of a gradient is

$$f(x) \approx f(x_0) + \partial f|_{x_0}[x - x_0],$$

where the approximation is "accurate " when $x \approx x_0$ and the linear function

$$\partial f|_{x_0}\colon \mathcal{X} \to \mathbb{R}$$

is the "gradient" of f at x_0. For this formulation to make sense, $\mathcal{X}$ needs to be a vector space so that $x - x_0$ is defined, $\mathcal{X}$ needs to have a nice dual space so that $\partial f|_{x_0} \in \mathcal{X}^*$, and f needs to be "differentiable". Banach spaces meets these requirements. However, the ODE $\dot{x}_t = -\nabla f(x_t)$ does not make sense in Banach spaces as we would expect

$$\dot{x}_t = \lim_{h\to 0} \frac{1}{h}(x_{t+h} - x_t) \in \mathcal{X}$$

but

$$\nabla f(x_t) \in \mathcal{X}^*.$$

To define and analyze gradient descent and gradient flow in Banach spaces, the "mirror descent" formulation provides one resolution. However, we shall take a different route.

Remarkably, gradient flow in metric spaces can be defined through the following variational formulation. For $\varepsilon > 0$, let

$$\tilde{x}_{t+\varepsilon} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ f(x) + \frac{1}{2\varepsilon} d(x, \tilde{x}_t)^2 \right\},$$

where $f\colon \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ and $\mathcal{X}$ is a metric space. We say $\{x_t\}_{t\geq 0}$ is a gradient flow with respect to f with starting point x_0 if

$$\{\tilde{x}_{\lfloor t/\varepsilon \rfloor \varepsilon}\}_{t\geq 0} \to \{x_t\}_{t\geq 0}$$

in the topology of uniform convergence on compacta for some sequence $\{\varepsilon_k\}_{k\in\mathbb{N}} \in \mathbb{R}_{++}$ such that $\varepsilon_k \to 0$.

Although existence and uniqueness of such gradient flows is not always guaranteed, even for Banach spaces, we do have existence and uniqueness for the Wasserstein gradient flow that we consider. (Because the loss functions we consider are λ-semiconvex.)

7.4.2 Prelminaries: First variation

Let $\Theta \subset \mathbb{R}^d$ be nonempty. Consider $\mathcal{P}(\Theta) \subseteq \mathcal{M}_+(\Theta)$. (So $\mathcal{P}(\Theta)$ is not a vector space.) Let $\mathcal{L} : \mathcal{P}(\Theta) \to \mathbb{R}$. We call

$$\frac{\delta \mathcal{L}}{\delta \rho}\Big|_\rho \in \mathcal{C}(\Theta)$$

a *first variation* of $\mathcal{L}$ at ρ if

$$\frac{d}{dh}\mathcal{L}(\rho + h\nu)\Big|_{h=0^+} = \left\langle \frac{\delta \mathcal{L}}{\delta \rho}\Big|_\rho, \nu \right\rangle = \int_\Theta \frac{\delta \mathcal{L}}{\delta \rho}\Big|_\rho(\theta)\, d\nu(\theta)$$

for all ν such that $\mu + h\nu \in \mathcal{P}$ for sufficiently small $h > 0$.

The definition of a first variation is similar to, but not the same as, the Fréchet derivative. Precisely, $\frac{\delta \mathcal{L}}{\delta \rho}$ is not a Fréchet derivative as $\mathcal{P}(\Theta)$ is not a vector space. We are also requiring that the first variation is in $\mathcal{C}(\Theta)$, rather than $\mathcal{M}^*(\Theta)$, as would be required by a Fréchet derivative. We impose this requirement because $\mathcal{M}^*(\Theta)$ is a very difficult space to work with. The first variation may or may not exist, and it is not unique. In our setup, we will only consider nice functions for which the first variation exists, and it will be unique only up to a constant, as the admissible variations ν must be a signed measure with total mass 0. So

$$\begin{aligned}\mathcal{L}(\rho + h\nu) &\approx \mathcal{L}(\rho) + h \int_\Theta \frac{\delta \mathcal{L}}{\delta \rho}\Big|_\rho(\theta)\, d\nu(\theta) \\ &= \mathcal{L}(\rho) + h \int_\Theta \left(\frac{\delta \mathcal{L}}{\delta \rho}\Big|_\rho(\theta) + C \right) d\nu(\theta)\end{aligned}$$

for any constant $C \in \mathbb{R}$.

Theorem 45. *Assume* $\mathrm{supp}(\mu) = \Theta$. *Write* τ_{opt} *for the solution of the dual problem for* $W_2(\mu, \nu)$. *(Although we did not prove this, a solution exists.) Then we have the first-variation*

$$\frac{\delta}{\delta \mu} W_2(\mu, \nu)^2 = \frac{1}{2} \| \cdot \|^2 - \tau_{\mathrm{opt}}.$$

(Although τ_{opt} *is not unique, the optimal transport plan* $T = \nabla \tau_{\mathrm{opt}} \colon \Theta \to \Phi$ *is unique except on a* μ*-null set.)*

Proof outline. The proof roughly relies on the principle of the envelope theorem. Assume everything is differentiable. If

$$\begin{aligned} V(\alpha) &= \sup_{\beta} U(\alpha, \beta) \\ &= U(\alpha, \beta(\alpha)) \end{aligned}$$

where $\beta^\star(\alpha) \in \mathrm{argmax}_\beta \, U(\alpha, \beta)$. We assume the maximizer exists. Then,

$$\begin{aligned} \frac{d}{d\alpha} V(\alpha) &= \frac{d}{d\alpha} U(\alpha, \beta^\star(\alpha)) \\ &= \frac{\partial}{\partial \alpha} U(\alpha, \beta^\star(\alpha)) + \underbrace{\frac{\partial}{\partial \beta} U(\alpha, \beta^\star(\alpha))}_{=0} \frac{d\beta^\star(\alpha)}{d\alpha} \\ &= \frac{\partial}{\partial \alpha} U(\alpha, \beta^\star(\alpha)). \end{aligned}$$

Since

$$W_2^2(\mu, \nu) = \sup_{\tau} \left\{ \int \frac{1}{2} \| \cdot \|^2 - \tau d\mu + \int \frac{1}{2} \| \cdot \|^2 - \tau^* d\nu \right\},$$

we conclude the statement. □

7.4.3 Wasserstein gradient flow

The Wasserstein gradient flow is defined through

$$\tilde{\rho}_{t+\varepsilon} = \operatorname*{argmin}_{\rho \in \mathcal{P}(\Theta)} \left\{ \mathcal{L}(\rho) + \frac{1}{\varepsilon} W_2(\rho, \tilde{\rho}_t)^2 \right\}.$$

We assume $\mathcal{L}$ has a first variation. For now, assume $\mathrm{supp}(\rho_0) = \mathrm{supp}(\rho_t) = \Theta$. Then

$$C = \frac{\delta \mathcal{L}}{\delta \rho}\Big|_{\rho_{t+\varepsilon}} + \frac{1}{\varepsilon} \left(\frac{1}{2} \| \cdot \|^2 - \tau_{\mathrm{opt}} \right),$$

for some constant C, i.e.,

$$C = \frac{\delta \mathcal{L}}{\delta \rho}\Big|_{\rho_{t+\varepsilon}}(\theta) + \frac{1}{\varepsilon}\left(\frac{1}{2}\|\theta\|^2 - \tau_{\text{opt}}(\theta)\right), \qquad \forall\, \theta \in \Theta.$$

This follows from

$$\begin{aligned} 0 &= \frac{d}{dh}\mathcal{L}(\rho_{t+\varepsilon} + h\nu) + \frac{1}{2\varepsilon}W_2^2(\rho_{t+\varepsilon} + h\nu, \rho_t)\Big|_{h=0} \\ &= \left\langle \frac{\delta \mathcal{L}}{\delta \rho}\Big|_{\rho_{t+\varepsilon}} + \frac{1}{\varepsilon}\left(\frac{1}{2}\|\cdot\|^2 - \tau_{\text{opt}}\right), \nu \right\rangle \end{aligned}$$

for all $\nu \in \mathcal{P}(\Theta)$ such that $\rho_{t+\varepsilon} + h\nu \in \mathcal{P}(\Theta)$ for small enough $h > 0$, i.e. ν is an admissible perturbation. Since ν has zero total mass, the vanishing directional (Gateaux) derivative means the first variation is a constant. We further take the gradient of the first variation to get

$$0 = \nabla_\theta \frac{\delta \mathcal{L}}{\delta \rho}\Big|_{\rho_{t+\varepsilon}} + \frac{1}{\varepsilon}(I - T),$$

i.e.,

$$0 = \nabla_\theta \frac{\delta \mathcal{L}}{\delta \rho}\Big|_{\rho_{t+\varepsilon}}(\theta) + \frac{1}{\varepsilon}(\theta - T(\theta)), \qquad \forall\, \theta \in \Theta,$$

where $T = \nabla \tau_{\text{opt}}$ is the optimal transport plan from $\rho_{t+\varepsilon}$ to ρ_t (not the other way around). So

$$\mathbf{v}_t(\theta) = \frac{1}{\varepsilon}(\theta - T(\theta))$$

is the negative displacement $(t+\varepsilon, \theta) \mapsto (t, T(\theta))$, or, equivalently, the positive displacement $(t, T(\theta)) \mapsto (t+\varepsilon, \theta)$. (Mind the sign.) So as $\varepsilon \to 0$, $\mathbf{v}_t(\theta)$ becomes velocity of particles at (t, θ). Therefore,

$$\partial_t \rho_t + \operatorname{div}(\mathbf{v}_t \rho_t) = 0 \qquad \mathbf{v}_t = -\nabla \frac{\delta \mathcal{L}}{\delta \rho}.$$

Bibliographical notes

A good reference is: Geometric Flows for Applied Mathematicians Xiaohui Chen `http://publish.illinois.edu/xiaohuichen/files/2020/12/geometric_flows.pdf`

Also, Lenaïc Chizat's course notes. `https://lchizat.github.io/ot2021orsay.html`

Also { Euclidean, Metric, and Wasserstein } Gradient Flows: an overview Filippo Santambrogio `https://arxiv.org/abs/1609.03890`

Chapter 8

Mean-field theory

We are now ready to talk about the mean-field theory, which considers the population dynamics of the infinitely many neurons. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty. Consider the setup with training data $X \in \mathcal{X}$ and corresponding labels $Y \in \mathcal{Y}$. For the sake of concreteness and simplicity, let $\mathcal{Y} = \mathbb{R}$ and $Y = f_\star(X)$ for some true unknown $f_\star$. We focus on the square loss function, although we do set up the notation to allow for further generality. Let $P \in \mathcal{P}(\mathcal{X})$ be a probability measure on the data.

Consider the risk function $R\colon L^2(P) \to \mathbb{R}_+$ defined as

$$R[f] = \mathbb{E}_{X\sim P}\frac{1}{2}\|f(X) - f_\star(X)\|^2 = \frac{1}{2}\langle f - f_\star, f - f_\star\rangle_{L^2(P)}.$$

Note, $\partial_f R|_{f_0} = f_0 - f_\star$.

Let M be the number of neurons. Let $\theta_i = (u_i, a_i, b_i) \in \mathbb{R}^{d+2}$ be parameters, and we collectively denote them by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$. Let $\Phi\colon \mathbb{R}^d \times \mathbb{R}^{d+2} \to \mathbb{R}$ be defined as

$$\Phi(x; \theta_i) = u_i \sigma(a_i^T x + b_i)$$

for some $\sigma\colon \mathbb{R} \to \mathbb{R}$. Then the M-wide 2-layer MLP $f_\theta^{(M)}$ is defined by

$$f_{\boldsymbol{\theta}}^{(M)}(x) = \frac{1}{M}\sum_{i=1}^{M} u_i \sigma(a_i^T x + b_i) = \frac{1}{M}\sum_{i=1}^{M} \Phi(x; \theta_i).$$

Now, consider training through

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad R[f_{\boldsymbol{\theta}}^{(M)}].$$

8.1 Convergence of risk

The parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)$ are initialized as IID samples from $\rho_0 \in \mathcal{P}(\Theta)$. The neural network is rewritten as

$$f_{\boldsymbol{\theta}}^{(M)} = \int \Phi(\cdot;\theta) d\rho_0^{(M)}(\theta), \qquad \rho_0^{(M)} = \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta_i}.$$

(By LLN, $\rho_0^{(M)} \rightharpoonup \rho_0$ a.s. as $M \to \infty$.) For notational simplicity, write

$$\int \Phi(\cdot;\theta) d\rho_0^{(M)}(\theta) = \int \Phi d\rho_0^{(M)}$$

Then,

$$R\left[f_{\boldsymbol{\theta}}^{(M)}\right] \to R\left[\int \Phi d\rho_0\right]$$

by the law of large numbers. Then, $f_{\boldsymbol{\theta}}^{(M)}(x)$ converges to $\int \Phi(x;\theta) d\rho_0(\theta)$ pointwise.

To see why, note

$$\begin{aligned}
R\left[f_{\boldsymbol{\theta}}^{(M)}\right] &= \frac{1}{2}\mathbb{E}_{X\sim P}\left[\int \Phi(X;\theta) d\rho_0^{(M)}(\theta) \int \Phi(X;\theta') d\rho_0^{(M)}(\theta')\right] \\
&\qquad - \mathbb{E}_{X\sim P}\left[\int \Phi(X;\theta) d\rho_0^{(M)}(\theta) f_\star(X)\right] + \frac{1}{2}\mathbb{E}_{X\sim P}\left[f_\star(X)^2\right] \\
&= \frac{1}{2}\int\int \underbrace{\mathbb{E}_{X\sim P}\left[\Phi(X;\theta)\Phi(X;\theta')\right]}_{=U(\theta,\theta')} d\rho_0^{(M)}(\theta) d\rho_0^{(M)}(\theta') \\
&\qquad - \int \underbrace{\mathbb{E}_{X\sim P}\left[\Phi(X;\theta) f_\star(X)\right]}_{=V(\theta)} d\rho_0^{(M)}(\theta) + \frac{1}{2}\mathbb{E}_{X\sim P}\left[f_\star(X)^2\right] \\
&\to R\left[\int \Phi d\rho_0\right]
\end{aligned}$$

where the second equality follows from Fubini.

Define

$$\begin{aligned}
U(\theta,\theta') &= \mathbb{E}_{X\sim P}\left[\Phi(X;\theta)\Phi(X;\theta')\right] \\
V(\theta) &= \mathbb{E}_{X\sim P}\left[\Phi(X;\theta) f_\star(X)\right].
\end{aligned}$$

(Then $U\colon \mathbb{R}^{d+2} \times \mathbb{R}^{d+2} \to \mathbb{R}$ is a PDK.) Define $\tilde{R}\colon \mathcal{P}(\mathbb{R}^{d+2}) \to \mathbb{R}$ as

$$\tilde{R}(\rho) = R\left[\int \Phi d\rho\right] = \int_{\mathbb{R}^{d+2}}\int_{\mathbb{R}^{d+2}} U(\theta,\theta')\, d\rho(\theta) d\rho(\theta') - \int_{\mathbb{R}^{d+2}} V(\theta)\, d\rho(\theta) + C.$$

where C is a constant independent of ρ. Then

$$R\left[f_{\boldsymbol{\theta}}^{(M)}\right] = \tilde{R}(\rho_0^{(M)}).$$

It is straightforward to show that $\frac{\delta \tilde{R}}{\delta \rho}\Big|_\rho(\cdot) = \int U(\cdot, \theta')\ d\rho(\theta') - V(\cdot)$. (Cf. Homework exercise.)

8.2 Population dynamics from gradient flow

Derivation for quadratic loss. We first derive the result for the setup with quadratic loss. We start with $M < \infty$. Let

$$\rho_t^{(M)} = \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta_i(t)}$$

with the parameters governed by gradient flow

$$\dot{\boldsymbol{\theta}} = -M\nabla_{\boldsymbol{\theta}} R[f_{\boldsymbol{\theta}}]$$

$$\begin{aligned}
\dot{\theta}_i &= -M\nabla_{\theta_i} R\left[f_{\boldsymbol{\theta}}^{(M)}\right] = -M\mathbb{E}_{X\sim P}\left[\left(f_{\boldsymbol{\theta}}^{(M)} - f_\star(x)\right)\nabla_{\theta_i} f_{\boldsymbol{\theta}}^{(M)}(x)\right] \\
&= -\mathbb{E}_{X\sim P}\left[\left(\int \Phi(x;\theta)\ d\rho_t^{(M)}(\theta) - f_\star(x)\right)\nabla_{\theta_i}\Phi(x;\theta_i)\right] \\
&= -\int \mathbb{E}_X\left[\Phi(x;\theta)\nabla_{\theta_i}\Phi(x;\theta_i)\right] d\rho_t^{(M)}(\theta) + \mathbb{E}_X\left[\nabla_{\theta_i}\Phi(x;\theta_i) f_\star(x)\right] \\
&= -\nabla_{\theta_i}\left(\int U(\theta_i,\theta') d\rho_t^{(M)}(\theta')\right) + \nabla_{\theta_i} V(\theta_i) \\
&= -\nabla_{\theta_i}\frac{\delta \tilde{R}}{\delta \rho}\Big|_{\rho_t^{(M)}}(\theta_i)
\end{aligned}$$

This result describes the velocity of the ith particle at position θ_i. Therefore, the population dynamics (of the M neurons) is described by the continuity equation

$$\partial_t \rho_t^{(M)} = \operatorname{div}\left(\rho_t^{(M)}\nabla\frac{\delta \tilde{R}}{\delta \rho}\Big|_{\rho_t^{(M)}}\right).$$

We will more rigorously justify this soon in Lemma 21.

General derivation for general loss. We now derive the result in the further general setup:

$$\begin{aligned}\dot{\theta}_i &= -M\nabla_{\theta_i} R\left[f_\theta^{(M)}\right] = -M\nabla_{\theta_i} R\left[\frac{1}{M}\sum_{j=1}^{M}\Phi(\cdot;\theta_j)\right] \\ &= -\left\langle \partial_f R|_{\frac{1}{M}\sum_{j=1}^{M}\Phi(\cdot;\theta_j)}(\bullet), \nabla_{\theta_i}\Phi(\bullet;\theta_i)\right\rangle_{L^2(P)} \\ &= -\nabla_{\theta_i}\langle \partial_f R, \Phi(\cdot;\theta_i)\rangle_{L^2(P)} \\ &= -\nabla_{\theta_i}\frac{\delta\tilde{R}}{\delta\rho}\bigg|_{\rho_t^{(M)}}(\theta_i)\end{aligned}$$

where the second line follows from a chain-rule type of argument. (Cf. Homework assignment.) For notational simplicity, define the *mean potential*

$$J(\theta|\rho) = \frac{\delta\tilde{R}}{\delta\rho}\bigg|_{\rho}(\theta) = \left\langle \Phi(\bullet;\theta), \partial_f R|_{\int_{\mathbb{R}^{d+2}}\Phi(\cdot;\theta')\,d\rho(\theta')}(\bullet)\right\rangle_{L^2(P)}.$$

Then

$$\dot{\theta}_i = -\nabla_{\theta_i} J(\theta_i|\rho_t^{(M)})$$

and the population dynamics is governed by the continuity equation:

$$\partial_t \rho_t^{(M)}(\theta) = \mathrm{div}(\rho_t^{(M)}(\theta)\nabla_\theta J(\theta|\rho_t^{(M)}))$$

or, more concisely,

$$\partial_t \rho_t^{(M)} = \mathrm{div}(\rho_t^{(M)}\nabla J(\cdot|\rho_t^{(M)})).$$

Let us prove this rigorously.

Lemma 21. *Let $\boldsymbol{\theta}(t) = (\theta_1(t), \ldots, \theta_M(t))$. The dynamics*

$$\dot{\theta}_i = \mathbf{v}_t(\theta_i; \rho_t^{(M)}), \qquad \forall\, i = 1, \ldots, M$$

satisfies the continuity equation

$$\partial_t \rho_t^{(M)} = -\mathrm{div}(\rho_t^{(M)}\mathbf{v}_t(\cdot; \rho_t^{(M)})).$$

Proof. For any $\varphi_t(\theta) \in \mathcal{C}_c^\infty\left((0,\infty) \times \mathbb{R}^{d+2}\right)$,

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^{d+2}} \left(\partial_t \varphi_t(\theta) + \nabla_\theta \varphi_t(\theta) \cdot \mathbf{v}_t\right) d\rho_t^{(M)}(\theta) dt \\
&= \frac{1}{M} \sum_{i=1}^M \int_0^\infty \partial_t \varphi_t(\theta_i(t)) + \nabla_\theta \varphi_t(\theta_i(t)) \dot{\theta}_i(t) dt \\
&= \frac{1}{M} \sum_{i=1}^M \int_0^\infty \frac{d}{dt} \left(\varphi_t(\theta_i(t))\right) dt \\
&= \frac{1}{M} \sum_{i=1}^M \left(\varphi_{t=\infty}(\cdot) - \varphi_{t=0}(\cdot)\right) \\
&= 0.
\end{aligned}$$

□

Remember that $\rho_0^{(M)} \rightharpoonup \rho_0$ as $M \to \infty$. Define ρ_t to be the solution of

$$\partial_t \rho_t = \operatorname{div}(\rho_t \nabla J(\cdot|\rho_t))$$

with initial condition ρ_0 at time $t = 0$. Then, for any $t \geq 0$, $\rho_t^{(M)} \rightharpoonup \rho_t$ as $M \to \infty$, although we do not prove this. The following (kind of) commutative diagram captures this idea:

$$\begin{array}{ccc}
\rho_0^{(M)} & \xrightarrow{(1)} & \rho_0 \\
\Big\downarrow{\scriptstyle (2)} & & \Big\downarrow{\scriptstyle (3)} \\
\rho_t^{(M)} & \xrightarrow{(4)} & \rho_t
\end{array}$$

(1) Follows from the law of large numbers.

(2) Denotes the time evolution of $\rho_t^{(M)}$ induced by gradient flow, which is equivalent to the continuity equation by Lemma 21.

(3) Denotes the definition of ρ_t via the continuity equation.

(4) Follows from the fact that the PDE is well posed and thus the time evolution of the continuity equation is continuous (in the weak topology) with respect to the initial condition, i.e., $\rho_0^{(M)} \rightharpoonup \rho_0$ implies $\rho_t^{(M)} \rightharpoonup \rho_t$,.

If we can show that ρ_t "converges" to a desired result, we can conclude the parameter population of the finite neural network $\rho_0^{(M)}$ will also "converge" to the same result, if M is sufficiently large. We analyze the dynamics of ρ_t.

8.3 Global convergence

Throughout this section, assume ρ_t is the solution of

$$\partial_t \rho_t = \operatorname{div}(\rho_t \nabla J(\cdot|\rho_t)).$$

with initial condition ρ_0. Assume for simplicity that ρ_0 has full support, i.e., that $\operatorname{supp}(\rho_0) = \mathbb{R}^{d+2}$. Then a weak solution to the continuity equation uniquely exists, although we do not show this.

Lemma 22. *$\tilde{R}(\rho_t)$ is a nonincreasing function of t.*

Proof outline. The variational characterization of $\{\rho_t\}_{t\geq 0}$ via Wasserstein gradient flow leads to

$$\tilde{R}(\tilde{\rho}_{t+\varepsilon}) + \frac{1}{\varepsilon} W_2^2(\tilde{\rho}_{t+\varepsilon}, \tilde{\rho}_t) \leq \tilde{R}(\tilde{\rho}_t)$$

Hence,

$$\tilde{R}(\tilde{\rho}_{t+\varepsilon}) \leq \tilde{R}(\tilde{\rho}_t)$$

and with some additional arguments, we conclude

$$\tilde{R}(\rho_{t+\varepsilon}) \leq \tilde{R}(\rho_t).$$

□

Since $\{\tilde{R}(\rho_t)\}_{t\geq 0}$ is a nonincreasing real-valued sequence, it will converge. Without further assumptions, however, this does not imply that $\{\rho_t\}_{t\geq 0}$ (weakly) converges. Therefore, we will assume the convergence of the measures.

Theorem 46. *Assume $\rho_t \rightharpoonup \rho_\infty$. Then $\nabla J(\theta, \rho_\infty) = 0$ for all $\theta \in \operatorname{supp}(\rho_\infty)$*

Proof. By continuity arguments, we have that

$$0 = \partial_t \rho_\infty = \operatorname{div}(\rho_\infty \nabla J(\cdot|\rho_\infty)).$$

Formally using integration by parts (even though $J(\theta|\rho_\infty)$ is not smooth and compactly supported), we get

$$\begin{aligned} 0 &= \int_{\mathbb{R}^{d+2}} J(\theta|\rho_\infty) \operatorname{div}(\rho_\infty(\theta) \nabla J(\theta|\rho_\infty))\, d\theta \\ &= \int_{\mathbb{R}^{d+2}} J(\theta|\rho_\infty) \operatorname{div}(\nabla J(\theta|\rho_\infty))\, d\rho_\infty(\theta) \\ &= -\int_{\mathbb{R}^{d+2}} \|\nabla J(\theta|\rho_\infty)\|^2\, d\rho_\infty(\theta). \end{aligned}$$

and thus, $\nabla J(\theta|\rho_\infty) = 0$ ρ_∞-almost everywhere. □

As we will discuss soon, $\text{supp}(\rho_t) = \mathbb{R}^{d+2}$, but ρ_∞ may have a smaller support. However, the stationarity condition on ρ_∞ does not imply global optimality.

Since $\tilde{R}(\rho)$ is a convex function of ρ (for the losses we consider), all local minima of $\tilde{R}$ are global, and one may expect the gradient flow to converge to global minimum. This is true for the gradient flow associated with the total variation metric. However this is not true for the Wasserstein gradient flow. While there exists a notion of convexity for Wasserstein gradient flows, namely geodesic convexity, but $\tilde{R}$ is not geodesically convex in our setup.

Lemma 23. *Let $\tilde{R} : \mathcal{M}_+(\Theta) \to \mathbb{R}$ and let $\rho_\star \in \mathcal{M}_+(\Theta)$. Assume $\frac{\delta \tilde{R}}{\delta \rho}|_{\rho_\star}(\cdot) = J(\cdot|\rho_\star)$ exists. Then $\rho_\star$ minimizes (globally) $\tilde{R}$ if and only if*

$$J(\theta, \rho_\star) = 0 \text{ for } \theta \in \text{supp}(\rho_\star)$$
$$J(\theta, \rho_\star) \geq 0 \text{ for } \theta \notin \text{supp}(\rho_\star)$$

Proof outline. This is an infinite-dimensional version of the homework problem. □

Lemma 24. *Let $\tilde{R}\colon \mathcal{P}(\Theta) \to \mathbb{R}$ and let $\rho_\star \in \mathcal{P}(\Theta)$. Assume $\frac{\delta \tilde{R}}{\delta \rho}|_{\rho_\star}(\cdot) = J(\cdot|\rho_\star)$ exists. Then $\rho_\star$ minimizes (globally) $\tilde{R}$ if and only if*

$$J(\theta, \rho_\star) = c \text{ for } \theta \in \text{supp}(\rho_\star)$$
$$J(\theta, \rho_\star) \geq c \text{ for } \theta \notin \text{supp}(\rho_\star)$$

Proof outline. This is an infinite-dimensional version of the homework problem. □

To establish global convergence, we need further assumptions. We will utilize the homogeneity of the ReLU activation function.

Let σ be the ReLU activation function. This makes $\Phi(x; \theta) = u_i \sigma(a_i^T x + b_i)$ is nonnegative 2-homogeneous in θ. We can use homogeneity by reparameterizing each particle θ_i in polar coordinates as

$$\theta_i = r_i \eta_i, \text{ with } r_i \in \mathbb{R} \text{ and } \eta_i \in \mathcal{S}^{d+1}.$$

Using 2-homogeneity, we have

$$f_\theta^{(M)}(x) = \frac{1}{M} \sum_{i=1}^{M} \Phi(x; \theta_i) = \frac{1}{M} \sum_{i=1}^{M} r_i^2 \Phi(x; \eta_i).$$

Then, probability measure $\rho^{(M)} = \frac{1}{M}\sum_{i=1}^{M}\delta_{\theta_i} \in \mathcal{P}(\mathbb{R}^{d+2})$ then corresponds to the nonnegative measure

$$\nu^{(M)} = \frac{1}{M}\sum_{i=1}^{M} r_i^2 \delta_{\eta_i} \in \mathcal{M}_+(\mathcal{S}^{d+1})$$

in the sense that

$$f_\theta^{(M)}(x) = \int_{\mathbb{R}^{d+2}} \Phi(x;\theta)\, d\rho^{(M)}(\theta) = \int_{\mathcal{S}^{d+1}} \Phi(x;\eta)\, d\nu^{(M)}(\eta)$$

or more concisely

$$f_\theta^{(M)} = \int_{\mathbb{R}^{d+2}} \Phi\, d\rho^{(M)} = \int_{\mathcal{S}^{d+1}} \Phi\, d\nu^{(M)}.$$

More generally, given $\rho \in \mathcal{P}(\mathbb{R}^{d+2})$, define $\nu \in \mathcal{M}_+(\mathcal{S}^{d+1})$ via

$$\int_{\mathcal{S}^{d+1}} h(\eta)\, d\nu(\eta) = \int_{\mathbb{R}^{d+2}\setminus\{0\}} \|\theta\|^2 h(\theta/\|\theta\|)\, d\rho(\theta) \tag{8.1}$$

for all bounded measurable $h\colon \mathcal{S}^{d+1} \to \mathbb{R}$. Then

$$\int_{\mathbb{R}^{d+2}} \Phi\, d\rho = \int_{\mathcal{S}^{d+1}} \Phi\, d\nu.$$

Note that by 2-homogeneity of Φ,

$$J(\cdot|\rho) = J(\cdot|\nu) = \left\langle \Phi, \partial_f R|_{\int \Phi\, d\nu} \right\rangle_{L^2(P)}$$

with ρ and ν corresponding in the sense of (8.1).

The flow $\dot\theta_i = -\nabla_{\theta_i} J(\theta_i|\rho_t^{(M)})$, induces the dynamics:

$$\begin{cases} \dot r_i &= -2 r_i J(\eta_i|\nu_t) \\ \dot\eta_i &= -(I - \eta_i\eta_i^\intercal)\nabla J(\eta_i|\nu_t) \end{cases} \quad \text{with } \nu_t = \frac{1}{M}\sum_{i=1}^{M} r_i^2 \delta_{\eta_i}.$$

To see why,

$$\begin{aligned} r_i &= \|\theta_i\| \\ \dot r_i &= \frac{\langle \theta_i, \dot\theta_i\rangle}{\|\theta_i\|} \\ &= -\frac{2}{\|\theta_i\|} J(\theta_i|\rho_t^{(M)}) \\ &= -2 r_i J(\eta_i|\rho_t^{(M)}) \\ &= -2 r_i J(\eta_i|\nu_t^{(M)}). \end{aligned}$$

Likewise,

$$\begin{aligned}
\eta_i &= \frac{\theta_i}{\|\theta_i\|} \\
\dot{\eta}_i &= \frac{\dot{\theta}_i}{\|\theta_i\|} - \frac{\theta_i}{\|\theta_i\|^2}\frac{d}{dt}\|\theta_i\| = \frac{1}{\|\theta_i\|}\left(\dot{\theta}_i - \frac{\theta_i}{\|\theta_i\|^2}\langle\theta_i, \dot{\theta}_i\rangle\right) \\
&= \frac{1}{\|\theta_i\|}\left(-\nabla J(\theta_i|\rho^{(M)}) + \eta_i\eta_i^\intercal \nabla J(\theta_i|\rho^{(M)})\right) \\
&= -(I - \eta_i\eta_i^\intercal)\nabla J(\eta_i|\rho^{(M)}) \\
&= -(I - \eta_i\eta_i^\intercal)\nabla J(\eta_i|\nu^{(M)}),
\end{aligned}$$

where we use the fact that 2-homogeneity of $J(\cdot|\rho)$ implies 1-homogeneity of $\nabla J(\cdot|\rho)$ and the Euler identity.

We now derive the PDE describing the dynamics of ν. For now, consider $\nu_t^{(M)}$ corresponding to $\rho_t^{(M)}$ in the sense of (8.1). Consider a smooth test function $h\colon \mathbb{S}^{d+1} \to \mathbb{R}$. Then

$$\int_{\mathbb{S}^{d+1}} h(\eta)\, d\nu_t^{(M)}(\eta) = \frac{1}{M}\sum_{i=1}^M r_i^2 h(\eta_i),$$

and we have

$$\begin{aligned}
\frac{d}{dt}\int_{\mathbb{S}^{d+1}} h(\eta)\, d\nu_t^{(M)}(\eta) &= \frac{1}{M}\sum_{i=1}^M 2r_i\dot{r}_i h(\eta_i) + \frac{1}{M}\sum_{i=1}^M r_i^2 \nabla h(\eta_i)^\intercal \dot{\eta}_i \\
&= -\frac{1}{M}\sum_{i=1}^M 4r_i^2 J(\eta_i|\nu_t^{(M)})h(\eta_i) - \frac{1}{M}\sum_{i=1}^M r_i^2 \nabla h(\eta_i)^\intercal (I - \eta_i\eta_i^\intercal)\nabla J(\eta_i|\nu_t^{(M)}) \\
&= -4\int_{\mathbb{S}^{d+1}} h(\eta) J(\eta|\nu_t^{(M)})\, d\nu_t^{(M)}(\eta) - \int_{\mathbb{S}^{d+1}} \nabla h(\eta)^\intercal (I - \eta\eta^\intercal)\nabla J(\eta|\nu_t^{(M)})\, d\nu_t^{(M)}(\eta).
\end{aligned}$$

This yields the PDE

$$\partial_t \nu_t^{(M)}(\eta) = -4J(\eta|\nu_t^{(M)})\nu_t^{(M)}(\eta) + \operatorname{div}(\nu_t^{(M)}(\eta) P_\eta \nabla J(\eta|\nu_t^{(M)})) \tag{8.2}$$

where we use the Theorem 48 and div is the divergence on the Riemannian manifold $\mathcal{S}^{d+1}$ and $P_\eta = I - \eta\eta^\intercal$ is the projection onto the tangent space for $\eta \in \mathcal{S}^{d+1}$.

A similar argument can be carried out to obtain the PDE

$$\partial_t \nu_t(\eta) = -4J(\eta|\nu_t)\nu_t(\eta) + \operatorname{div}(\nu_t(\eta) P_\eta \nabla J(\eta|\nu_t))$$

for a general $\nu \in \mathcal{M}_+(\mathcal{S}^{d+1})$ corresponding to $\rho \in \mathcal{P}(\mathbb{R}^{d+2})$ in the sense of (8.1).

Theorem 47. *Assume the function $\Phi : \mathbb{S}^{d+1} \to \mathbb{R}$ is $(d+1)$-times continuously differentiable. Assume ν_0 is a nonnegative measure on the sphere $\mathbb{S}^{d+1}$ with finite mass and full support. Then the flow defined in PDE (8.2) is well defined for all $t \geq 0$. Moreover, if ν_t converges weakly to some limit ν_∞, then ν_∞ is a global minimum of the function $\nu \mapsto F(\nu) = R(\int_{\mathbb{S}^{d+1}} \Phi(\eta) d\nu(\eta))$ over the set of nonnegative measures.*

Proof. The existence and uniqueness of the flow $(\nu_t)_{t\leq 0}$ can be proved. By Lemma 24, to show that global minima ,we are enough to show that $J(\eta|\nu_\infty) = 0$ on the support of ν_∞ and $J(\eta|\nu_\infty) \geq 0$ on the entire sphere.

1. The support of ν_∞ The representation of the solution of PDE (8.2) as

$$\nu_t = X(t,\cdot)_\# \left(\nu_0 \exp\left(-4 \int_0^t J(X(s,\cdot)|\nu_s) ds \right) \right)$$

where $X : [0,\infty) \times \mathbb{S}^{d+1} \to \mathbb{S}^{d+1}$ is the flow associated to the time-dependent vector field $-P_\eta \nabla J(\cdot|\nu_t)$, i.e. it satisfies $X(0,\eta) = \eta$ and $\frac{d}{dt} X(t,\eta) = -P_\eta \nabla J(X(t,\eta)|\nu_t)$ for all $\eta \in \mathbb{S}^{d+1}$. Under regularity assumptions, some previous results for ODEs guarantee that $X(t,\cdot)$ is diffeomorphism of the sphere at all time t. Thus, image measure of ν_t is same as measure of the form $\nu_0 \exp(\cdot)$ which has full support. Thus v_t has full support.

2. global minimum We assume that the flow converges to some measure ν_∞. Then, $\partial_t \nu_t(\eta)|_{t=\infty} = 0 = -4J(\eta|\nu_\infty)\nu_\infty(\eta) + \nabla \cdot (\nu_\infty(\eta) P_\varphi \nabla J(\eta|\nu_\infty))$. Multiplying both side of the equation by J and integrate, then we get

$$\begin{aligned}
0 &= \int_{\mathbb{S}^{d+1}} J(\eta|\nu_\infty) \left\{ -4J(\eta|\nu_\infty) d\nu_\infty(\eta) + \nabla \cdot (\nu_\infty(\eta) P_\varphi \nabla J(\eta|\nu_\infty)) \right\} d\eta \\
&= \int_{\mathbb{S}^{d+1}} -4J^2(\eta|\nu_\infty) d\nu_\infty(\eta) + \int_{\mathbb{S}^{d+1}} J(\eta|\nu_\infty) \nabla \cdot (P_\varphi \nabla J(\eta|\nu_\infty)) d\nu_\infty(\eta) \\
&= -4 \int_{\mathbb{S}^{d+1}} J^2(\eta|\nu_\infty) d\nu_\infty(\eta) - \int_{\mathbb{S}^{d+1}} (P_\varphi \nabla J(\eta|\nu_\infty))^T (P_\varphi \nabla J(\eta|\nu_\infty)) d\nu_\infty(\eta)
\end{aligned}$$

Thus, J and $P_\varphi \nabla J$ is zero on the support of ν_∞ and no condition beyond the support of ν_∞. For contradiction, assume that $\inf_\eta J(\eta|\nu_\infty) < 0$. Then, by Sard's Theorem there is negative j such that $j > \int_\eta J(\eta|\nu_\infty)$ and the gradient $P_\eta \nabla J(\eta|\nu_\infty)$ does not vanish on the $\left\{\eta \in \mathbb{S}^{d+1} | J(\eta|\nu_\infty) = j\right\}$. we now consider the set $K = \{\eta | J(\eta|\nu_\infty \leq j\}$, which has some boundary ∂K, such that $P_\eta \nabla J(\eta|\nu_\infty) \cdot \hat{n} > 0$ where $\eta \in \partial K$ and outward normal vector $\hat{n}$.

Since ν_t converges weakly to ν_∞, there exists t_0 such that for all $t \geq t_0$ and $\eta \in K$, $J(\eta|\nu_t) < j/2$ and $P_\eta \nabla J(\eta|\nu_t) \cdot \hat{n} > 0$. Thus, for all $t > t_0$ and test

function $\varphi = 1_{\eta \in K}$,

$$\frac{d}{dt}\nu_t(K) = \frac{d}{dt}\int_K d\nu_t(\eta) = -4\int_K J(\eta|\nu_t)d\nu_t + \int_{\partial K} P_\eta \nabla J(\eta|\nu_t)\cdot \hat{n} d\nu_t$$
$$\geq -2j\nu_t(K).$$

Since $\nu_{t_0}(K) > 0$ (because ν_{t_0} has full support), $\nu_t(K)$ diverges, which is a contradiction with the convergence of ν_t. □

8.3.1 Differential geometry background

Let $\mathcal{M}$ be a compact Riemannian manifold[1] of finite dimension with empty boundary and g be the corresponding metric. (In this note, we consider only $\mathcal{M} = \mathbb{S}^{d+1}$ case.) Let $\varphi : \mathbb{S}^{d+1} \to \mathbb{R}$ be the test function and $V : \mathbb{S}^{d+1} \to \mathbb{R}^{d+2}$

1. $\varphi : \mathbb{S}^{d+1} \to \mathbb{R}$ is **smooth** if and only if there exists smooth function $\tilde{\varphi} : \mathbb{R}^{d+2} \to \mathbb{R}$ such that $\tilde{\varphi}|_{\mathcal{M}} = \varphi$.

2. $V = (V_1, V_2, \cdots, V_{d+2}) : \mathbb{S}^{d+1} \to \mathbb{R}^{d+2}$ is **smooth** if and only if V_i is smooth for all $i = 1, 2, \cdots, d+2$

3. Let p be a point of M. A linear map $v : \mathcal{C}^\infty(\mathcal{M}) \to \mathbb{R}$ is called a **derivation** at p if it is satisfies

$$v(fg) = f(p)vg + g(p)vf \text{ for all } f, g \in \mathcal{C}^\infty(\mathcal{M}).$$

4. The set of all derivations of $C^\infty(M)$ at p, denoted by $T_p(\mathcal{M})$, is a vector space called the **tangent space to** M **at** p. An element of $T_p\mathcal{M}$ is called a **tangent vector at** p**.**

5. (Existence of Local Orthonormal Frames). For each $p \in \mathcal{M}$, there is a smooth orthonormal frame $(E_1, \cdots, E_{d+1})$ on a neighborhood of p.

6. Then, the vectors $(E_1|_p, \cdot, E_{d+1}|_p)$ form an orthonormal basis for $T_p(\mathcal{M})$

7. $P_\eta : \mathbb{R}^{d+2} \to T_\eta$ is **orthogonal projection**. In linear algebra sense, $P_\eta = I - \eta\eta^T$ for $\eta \in \mathbb{S}^{d+1}$

8. Define **gradient** ∇f on $\mathbb{S}^{d+1}$ as a $\nabla f(\eta) = P_\eta(\nabla \tilde{f}(\eta)) \in T_\eta \subset \mathbb{R}^{d+2}$.

9. For $a, b \in T_\eta$, define **inner product** $\langle \cdot, \cdot \rangle_{T_\eta}$ as $\langle a, b \rangle_{T_\eta} = a^T b$.

[1]Riemannian manifold is a real, smooth manifold M equipped with a positive-definite inner product $\langle \cdot, \cdot \rangle_g$ on the tangent space at each point.

10. Define **divergence** of V as $\nabla \cdot V = \sum_{i=1}^{d+1} \frac{\partial \langle e_i, \tilde{V} \rangle}{\partial e_i}$ where $\tilde{V} : \mathbb{R}^{d+2} \to \mathbb{R}^{d+2}$ is smooth function such that $\tilde{V}|_{\mathbb{S}^{d+1}} = V$ and $e_1, \cdots, e_{d+1}$ are orthonormal basis for $T_\eta(\mathcal{M})$.

Theorem 48 (Integration by parts)**.** *Let $(\mathcal{M}, g)$ be a compact Riemannian manifold with boundary, let $\tilde{g}$ denote the induced Riemanninan metric on $\partial\mathcal{M}$, and let N be the outward unit normal vector field along $\partial\mathcal{M}$. Then, for $f \in C^\infty(\mathcal{M})$ and smooth vector field X, this satisfies* $\mathrm{div}(\varphi X) = \varphi \mathrm{div} X + \langle \mathrm{grad}\varphi, X \rangle_g$*. Furthermore, this satisfites the following "integration by parts" formula:*

$$\int_{\mathcal{M}} \langle \mathrm{grad}\varphi, X \rangle_g dV_g = \int_{\partial\mathcal{M}} \varphi \langle X, N \rangle_g dV_{\tilde{g}} - \int_{\mathcal{M}} (\varphi \mathrm{div} X) dV_g.$$

Then, Theorem 48 yields

$$\begin{aligned} \int_{\mathcal{M}} \nabla \cdot (\varphi V)\, d\eta &= \int_{\partial\mathcal{M}} \langle \varphi V, n \rangle\, dS = 0 \\ &= \int_{\mathcal{M}} \varphi \nabla \cdot V d\eta + \int_{\mathcal{M}} \langle \nabla\varphi, V \rangle_{T_\eta} d\eta. \end{aligned}$$

Consequently,

$$\begin{aligned} \int_{\mathbb{S}^{d+1}} \nabla\varphi(\eta)(I - \eta\eta^T)\nabla J(\eta|\nu)\, d\nu(\eta) &= \int_{\mathbb{S}^{d+1}} \langle \nabla\varphi(\eta), \nabla J(\eta|\nu) \rangle_{T_\eta}\, d\nu(\eta) \\ &= -\int_{\mathbb{S}^{d+1}} \varphi(\eta)\, (\nabla \cdot \nu(\eta) \nabla J(\eta|\nu))\, d\eta \end{aligned}$$

Lemma 25 (Sard's Theorem)**.** *Suppose M and N are smooth manifolds with or without boundary and $F : M \to N$ is a smooth map. Then the set of critical values of F has measure zero in N.*

Lemma 26 (Euler identity)**.** *If $F : \mathbb{R}^d \to \mathbb{R}$ is 2-homogeneous function, then $2F(\theta) = \theta^T \nabla F(\theta)$ for $\theta \in \mathbb{R}^d$.*

Chapter 9

Universal approximation theory: Deep neural networks

Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ and $d \in \mathbb{N}$. Define $\mathcal{NN}^{\sigma}_{d,m}$ be the class of MLPs with input $\mathbb{R}^d$, output $\mathbb{R}$, m intermediate neurons, and arbitrary finite depth. The activation function σ is applied after all layers, except the final layer. We will now show universality of $\mathcal{NN}^{\sigma}_{d,m}$ under a very mild assumption on σ.

Lemma 27. *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be a continuous function that is continuously differentiable at at least one point, with nonzero derivative at that point. Let $K \subseteq \mathbb{R}$ be compact. Then a neuron with activation function σ may uniformly approximate the identity function on K.*

Proof. Let r_0 be the point at which $\sigma'(r_0) \neq 0$ exists. Define $\tau_M\colon \mathbb{R} \to \mathbb{R}$ as

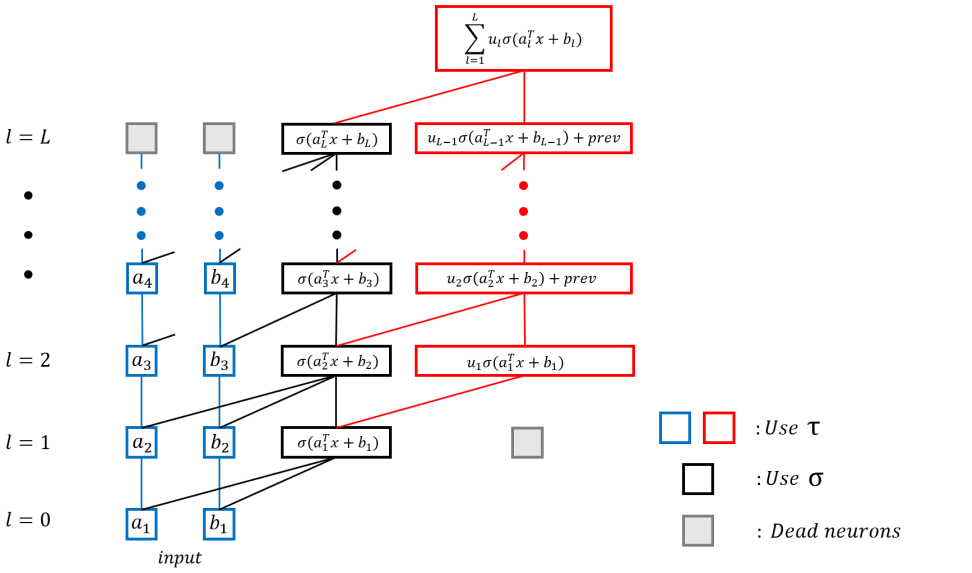
$$\tau_M(r) = \frac{1}{M\sigma'(r_0)} \left(\sigma(r_0 + Mr) - \sigma(r_0)\right).$$

As $M \to 0$, this uniformly approximates the identity function. □

Theorem 49. *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be continuous non-polynomial function which is continuously differentialble at at least one point, with nonzero derivative at that point. Let $\Omega \subset \mathbb{R}^d$ be compact. Then $\mathcal{NN}^{\sigma}_{d,d+2}$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.*

Proof. Define $\mathcal{NN}^{\sigma,\tau}_{d,m}$ be the class of MLPs similar to $\mathcal{NN}^{\sigma}_{d,m}$, except that we have the freedom to choose either σ or τ for the activation function. Let τ be the identity function.

By Lemma 27, we have $\overline{\mathcal{NN}^{\sigma,\tau}_{d+2}} = \overline{\mathcal{NN}^{\sigma}_{d+2}}$. Since σ is non-polynomial, the classical universal approximation theorem tells us that $\sum_{l=1}^{L} u_l \sigma(a_l^T x + b_l)$ is dense. Finally, consider the architecture:



□

Lemma 28. *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be any nonaffine polynomial. Let $K \subseteq \mathbb{R}$ be compact. Then a neuron with activation function σ many uniformly approximate the quadratic function $\kappa\colon x \mapsto x^2$ on K.*

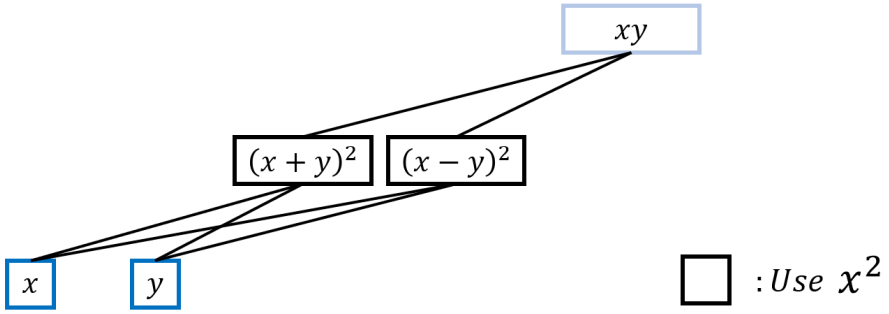
Proof. Fix $r_1 \in \mathbb{R}$ such that $\sigma''(r_1) \neq 0$, which exist as σ is nonaffine. Define $\kappa_M\colon \mathbb{R} \to \mathbb{R}$ by

$$\kappa_M(r) = \frac{\sigma(r_1 + Mr) - 2\sigma(r_1) + \sigma(r_1 - Mr)}{M^2\sigma''(r_1)}.$$

As $M \to 0$, this uniformly approximates the quadratic function κ. □

Lemma 29. *Using the square and the identity activation functions, the multiplication operation $(x, y) \mapsto xy$ for $x, y \in \mathbb{R}$ can be represented with 2 intermediate neurons and one additional neuron storing the output.*

Proof. Consider the architecture:



□

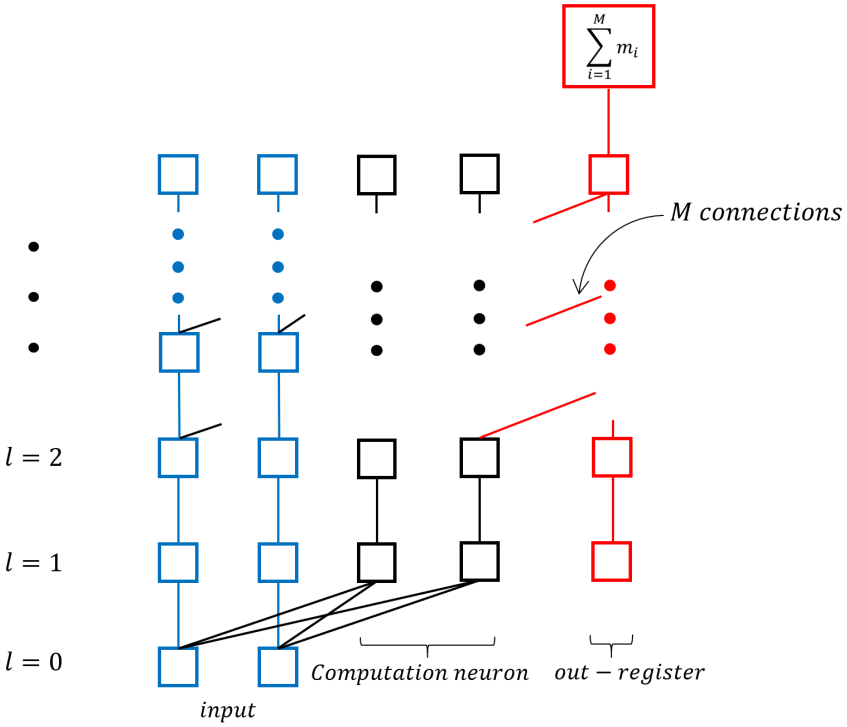
Theorem 50. *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be nonaffine polynomial. Let $\Omega \subset \mathbb{R}^d$ be compact. Then $\mathcal{NN}^{\sigma}_{d,d+3}$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$. (With a tighter analysis, the width $d+3$ can be reduced to $d+2$.)*

Proof. Define $\mathcal{NN}^{\kappa,\tau}_{d,m}$ be the class of MLPs similar to $\mathcal{NN}^{\sigma}_{d,m}$, except that we have the freedom to choose either κ or τ for the activation function. Let κ be the square function and τ be the identity function. By Lemmas 27 and 28, $\overline{\mathcal{NN}^{\kappa,\tau}_{d,d+3}} \subseteq \overline{\mathcal{NN}^{\sigma}_{d,d+3}}$. If we show that $\mathcal{NN}^{\kappa,\tau}_{d,d+3}$ contains any polynomial (of d variables) then we are done, since the algebra of polynomials is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$ by Stone–Weierstrass.

Consider any

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{M} \beta_i m_i, \quad m_i = \prod_{j=1}^{d} x_j^{\alpha_j}$$

for $i = 1, \cdots, M$ and $\alpha_i \in \mathbb{N}$. Consider the architecture:



We designate d neurons to copy the input and 2 neurons to perform the multiplication operation described in Lemma 29. The computation neurons carry out the products forming the monomials $m_1, \ldots, m_M$. Then the out-register

neurons accumulate the result of these M monomials by performing addition M times. $\square$

Chapter 10

Neural ODE

Consider the depth-L residual network

$$\begin{aligned} h_\theta(X) &= z_L \\ z_L &= z_{L-1} + f(z_{L-1}, \theta, L-1) \\ &\vdots \\ z_2 &= z_1 + f(z_1, \theta, 1) \\ z_1 &= z_0 + f(z_0, \theta, 0) \\ z_0 &= X \end{aligned}$$

where $z_0, \dots, z_L \in \mathbb{R}^D$, $\theta \in \mathbb{R}^P$, and $f\colon \mathbb{R}^D \times \mathbb{R}^P \times \mathbb{N} \to \mathbb{R}^D$. Note that θ is shared across all layers. Consider the loss function

$$\text{loss} = \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(X_i), Y_i).$$

For simplicity, assume $N = 1$ and write

$$\mathcal{L} = \ell(h_\theta(X), Y),$$

where $\mathcal{L}$ loosely denotes the output scalar loss value.

The *neural ODE* is a continuous-depth (or infinite-depth) analog:

$$\begin{aligned} h_\theta(X) &= z(1) \\ \dot{z}(s) &= f\left(z(s), \theta, s\right) \qquad \text{for } s \in [0,1] \\ z(0) &= X, \end{aligned}$$

where $z(s) \in \mathbb{R}^D$ for $s \in [0,1]$, $\theta \in \mathbb{R}^P$, and $f\colon \mathbb{R}^D \times \mathbb{R}^P \times [0,1] \to \mathbb{R}^D$. Assume f is continuous in (z, θ, s) and continuously differentiable in (z, θ). The idea is

that $f(z, \theta, s)$ is represented by a neural network. More precisely, $\{z(s)\}_{s\in[0,1]}$ is a solution to this ODE if

$$z(s) = X + \int_0^s f\left(z(s'), s', \theta\right)\, ds', \qquad s \in [0, 1].$$

For simplicity, assume that the ODE has a unique solution.[1]

We refer to $s \in [0, 1]$ as "pseudo-time" to distinguish it from "time"; pseudo-time corresponds to the progression of depth while the time corresponds to the progression of training iterations. In this lecture, we will not consider training iterations of the neural ODE, so time will not appear.

Generally, one considers the loss function

$$\text{loss} = \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(X_i), Y_i),$$

where $h_\theta(X_i)$ is the solution to the ODE at pseudo-time $s = 1$ with initial condition $z(0) = X_i$ at pseudo-time $s = 0$. For simplicity, assume $N = 1$ and write

$$\mathcal{L} = \ell(h_\theta(X), Y),$$

where $\mathcal{L}$ loosely denotes the output scalar loss value.

The 2018 neural ODE paper by Chen, Rubanova, Bettencourt, and Duvenaud ignited an exciting line of empirical and theoretical research. From the empirical side, neural ODEs have found many interesting applications with strong benchmark results. In fact, research on and using neural ODEs is primarily experimental rather than theoretical. From the theoretical side, neural ODEs can be viewed as an infinite-depth limit of the ResNet, but the neural ODE, by itself, does not provide any trainability guarantees. In this lecture, we will discuss how to perform the continuous-depth analog of backpropagation on neural ODEs. In practice, Neural ODEs are trained using SGD with gradients computed via the following approach.

As a warmup exercise, let us carry out backpropagation of the discrete-depth ResNet. Assume the forward pass has been performed, i.e., $z_1, \ldots, z_L$ have been sequentially computed and their values been stored in memory. For notational simplicity, denote

$$a_l = \frac{\partial \mathcal{L}}{\partial z_l} = \frac{\partial \mathcal{L}}{\partial z_L}\frac{\partial z_L}{\partial z_{L-1}} \cdots \frac{\partial z_{l+2}}{\partial z_{l+1}}\frac{\partial z_{l+1}}{\partial z_l}, \qquad l = 0, \ldots, L.$$

[1] In practice, f will be represented by a neural network with continuous activation functions, so it is reasonable to assume f is locally Lipschitz continuous. By the Picard–Lindelöf theorem, local Lipschitz continuity implies that a solution exists for $s \in [0, \varepsilon)$ for some small $\varepsilon > 0$, but there is no a priori guarantee that $z(s)$ does not blow up within $s \in [0, 1]$.

Then we have

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} &= \frac{\partial \mathcal{L}}{\partial z_L}\frac{\partial z_L}{\partial \theta} \\
&= \frac{\partial \mathcal{L}}{\partial z_L}\left(\frac{\partial f}{\partial \theta}(z_{L-1}, \theta, L-1) + \frac{\partial f}{\partial z_{L-1}}(z_{L-1}, \theta, L-1)\frac{\partial z_{L-1}}{\partial \theta} + \frac{\partial z_{L-1}}{\partial \theta}\right) \\
&= a_L \frac{\partial f}{\partial \theta}(z_{L-1}, \theta, L-1) + \frac{\partial \mathcal{L}}{\partial z_L}\frac{\partial z_L}{\partial z_{L-1}}\frac{\partial z_{L-1}}{\partial \theta} \\
&= a_L \frac{\partial f}{\partial \theta}(z_{L-1}, \theta, L-1) + a_{L-1}\frac{\partial f}{\partial \theta}(z_{L-2}, \theta, L-2) + a_{L-1}\frac{\partial z_{L-1}}{\partial z_{L-2}}\frac{\partial z_{L-2}}{\partial \theta} \\
&= \sum_{l=1}^{L} a_l \frac{\partial f}{\partial \theta}(z_{l-1}, \theta, l-1).
\end{aligned}$$

Next, we will obtain an analogous formula for the neural ODE.

10.1 Backpropagation for neural ODE

10.1.1 Warmup for continuous-depth backprop

The full derivation of the continuous-depth backprop will be carried out soon in Theorem 51. However, let us carry out a smaller computation as a warmup to familiarize ourselves with the key technique.

We first introduce the some machinery and notation. For $s, t \in [0, 1]$, define the flow operator (also called the time evolution operator) $\mathcal{F}^{s,t} \colon \mathbb{R}^D \to \mathbb{R}^D$ as

$$\begin{aligned}
\mathcal{F}^{s,t}(z) &= z(t) \\
\dot{z}(s') &= f\left(z(s'), \theta, s'\right) \qquad \text{for } s' \in [s, t] \\
z(s) &= z.
\end{aligned}$$

Then

$$z(1) = \mathcal{F}^{0,1}(X) = \mathcal{F}^{s,1}(\mathcal{F}^{0,s}(X))$$

for any $s \in [0, 1]$.

The flow operator can evolve the initial condition forward in pseudo-time ($t > s$) and also backwards in pseudo-time ($t < s$), since the ODE can be solved both forwards and backwards in pseudo-time. In fact, if $z(1)$ is known, then the initial condition $z(0) = \mathcal{F}^{1,0}(z(1))$ can be recovered through solving the ODE

$$\begin{aligned}
\dot{z}(s) &= f\left(z(s), \theta, s\right) \qquad \text{for } s \in [0, 1] \\
z(1) &\qquad \text{“initial” condition.}
\end{aligned}$$

This was not the case in the discrete-depth ResNet; knowledge of z_L does not necessarily allow one to recover z_{L-1} or z_0. For continuous-depth neural ODE, obtaining $z(0)$ from $z(1)$ is no more difficult than obtaining $z(1)$ from $z(0)$.

Define

$$\frac{\partial \mathcal{L}}{\partial z(s)} = D(\mathcal{L} \circ \mathcal{F}^{s,1})(z(s)) = \left.\frac{\partial \mathcal{L}(\mathcal{F}^{s,1}(z))}{\partial z}\right|_{z=z(s)}$$

and

$$\frac{\partial z(t)}{\partial z(s)} = D(\mathcal{F}^{s,t})(z(s)) = \left.\frac{\partial \mathcal{F}^{s,t}(z)}{\partial z}\right|_{z=z(s)}$$

for $s, t \in [0, 1]$. Then, we have the chain rule

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial z(s)} = D(\mathcal{L} \circ \mathcal{F}^{s,1})(z(s)) &= D(\mathcal{L} \circ \mathcal{F}^{t,1} \circ \mathcal{F}^{s,t})(z(s)) \\ &= D(\mathcal{L} \circ \mathcal{F}^{t,1})(z(t)) \cdot D(\mathcal{F}^{s,t})(z(s)) \\ &= \frac{\partial \mathcal{L}}{\partial z(t)} \frac{\partial z(t)}{\partial z(s)}\end{aligned}$$

for $s, t \in [0, 1]$. So $\frac{\partial \mathcal{L}}{\partial z(s)}$ represents the infinitesimal change in $\mathcal{L}$ if the neural ODE started at pseudo-time s with initial value $z(s) + \delta$, where δ is an infinitesimal perturbation.

Let

$$a(s) = \frac{\partial \mathcal{L}}{\partial z(s)} \in \mathbb{R}^{1 \times D}, \qquad s \in [0, 1].$$

Then

$$\begin{aligned}\dot{a}(s) &= -a(s) \frac{\partial f}{\partial z}(z(s), \theta, s), \qquad s \in [0, 1] \\ a(1) &= \frac{\partial \mathcal{L}}{\partial z(1)}\end{aligned}$$

and $\{a(s)\}_{s \in [0,1]}$ can be solved by solving the ODE backwards in pseudo-time with "initial condition" $a(1)$. (ODE solver is given $a(1)$ and solves for $a(s)$ for

$0 \leq s < 1$.) This follows from

$$\begin{aligned}
\dot{a}(s) &= \lim_{\varepsilon \to 0} \frac{a(s+\varepsilon) - a(s)}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{a(s+\varepsilon)}{\varepsilon} \left(I - \frac{\partial z(s+\varepsilon)}{\partial z(s)} \right) \\
&= \lim_{\varepsilon \to 0} \frac{a(s+\varepsilon)}{\varepsilon} \left(I - \frac{\partial}{\partial z(s)} \left(z(s) + \int_s^{s+\varepsilon} f(z(s'), \theta, s') \, ds' \right) \right) \\
&= - \lim_{\varepsilon \to 0} a(s+\varepsilon) \frac{\partial f(z(s), \theta, s)}{\partial z(s)} + \mathcal{O}(\varepsilon) \\
&= - \underbrace{a(s)}_{1 \times D} \underbrace{\frac{\partial f}{\partial z}(z(s), \theta, s)}_{D \times D} .
\end{aligned}$$

Ultimately, we want $\frac{\partial \mathcal{L}}{\partial \theta}$. However, infinitesimal changes of θ to $\theta + \delta$ affects the update as

$$z(s+\varepsilon) \approx z(s) + \varepsilon f(z(s), \theta + \delta, s)$$

and making sense of this precisely and correctly is tricky. Therefore, we employ another technique of converting θ into an initial condition of an augmented ODE.

10.1.2 Backprop via adjoint equations

Theorem 51. *Consider the neural ODE. The solution to the ODE*

$$\begin{aligned}
\dot{a}(s) &= -a(s) \frac{\partial f}{\partial z}(z(s), \theta, s), \quad \textit{for } s \in [0,1] \\
\dot{b}(s) &= -a(s) \frac{\partial f}{\partial \theta}(z(s), \theta, s), \quad \textit{for } s \in [0,1] \\
a(1) &= \frac{\partial \mathcal{L}}{\partial z(1)} \in \mathbb{R}^{1 \times D} \\
b(1) &= 0 \in \mathbb{R}^{1 \times P}
\end{aligned}$$

yields $\frac{\partial \mathcal{L}}{\partial \theta} = b(0)$.

Proof. Augment the ODE as follows:

$$\begin{aligned}
\dot{z}(s) &= f(z(s), \varphi(s), s), \quad \text{for } s \in [0,1] \\
\dot{\varphi}(s) &= 0, \quad \text{for } s \in [0,1] \\
z(0) &= X \\
\varphi(0) &= \theta.
\end{aligned}$$

Define the augmented notation

$$z_{\text{aug}}(s) = \begin{bmatrix} z(s) \\ \varphi(s) \end{bmatrix} \in \mathbb{R}^{(D+P)\times 1}$$
$$f_{\text{aug}}(z_{\text{aug}}(s), s) = \begin{bmatrix} f(z(s), \varphi(s), s) \\ 0 \end{bmatrix} \in \mathbb{R}^{(D+P)\times 1}.$$

Then

$$\begin{aligned} \dot{z}_{\text{aug}}(s) &= f_{\text{aug}}(z_{\text{aug}}(s), s), \qquad \text{for } s \in [0,1] \\ z_{\text{aug}}(0) &= \begin{bmatrix} X \\ \theta \end{bmatrix}. \end{aligned}$$

For $s, t \in [0,1]$, define the augmented flow operator $\mathcal{F}^{s,t}_{\text{aug}} \colon \mathbb{R}^{D+P} \to \mathbb{R}^{D+P}$ as

$$\begin{aligned} \mathcal{F}^{s,t}_{\text{aug}}(z, \varphi) &= (z(t), \varphi(t)) \\ \dot{z}_{\text{aug}}(s') &= f_{\text{aug}}(z_{\text{aug}}(s'), s'), \qquad \text{for } s' \in [s,t] \\ z_{\text{aug}}(s) &= \begin{bmatrix} z(s) \\ \varphi(s) \end{bmatrix}. \end{aligned}$$

Then define

$$\begin{aligned} a_{\text{aug}}(s) = \frac{\partial \mathcal{L}}{\partial z_{\text{aug}}(s)} &= \left.\frac{\partial \mathcal{L}(\mathcal{F}^{s,1}_{\text{aug}}(z_{\text{aug}}))}{\partial z_{\text{aug}}}\right|_{z_{\text{aug}} = z_{\text{aug}}(s)} \in \mathbb{R}^{1\times(D+P)} \\ &\overset{\text{(i)}}{=} \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial z(s)} & \frac{\partial \mathcal{L}}{\partial \varphi(s)} \end{bmatrix} \\ &\overset{\text{(ii)}}{=} \begin{bmatrix} a(s) & b(s) \end{bmatrix}, \end{aligned}$$

where (i) defines $\frac{\partial \mathcal{L}}{\partial z(s)}$ and $\frac{\partial \mathcal{L}}{\partial \varphi(s)}$ and (ii) defines $a(s)$ and $b(s)$. Alternatively, we can define

$$\begin{aligned} a(s) = \frac{\partial \mathcal{L}}{\partial z(s)} &= \left.\frac{\partial \mathcal{L}(\mathcal{F}^{s,1}_{\text{aug}}(z, \varphi))}{\partial z}\right|_{\substack{z = z(s) \\ \varphi = \varphi(s)}} \\ b(s) = \frac{\partial \mathcal{L}}{\partial \varphi(s)} &= \left.\frac{\partial \mathcal{L}(\mathcal{F}^{s,1}_{\text{aug}}(z, \varphi))}{\partial \varphi}\right|_{\substack{z = z(s) \\ \varphi = \varphi(s)}}. \end{aligned}$$

The meaning of $\frac{\partial \mathcal{L}}{\partial z(s)}$ remains essentially unchanged. The meaning of $\frac{\partial \mathcal{L}}{\partial \varphi(s)}$ is the infinitesimal change in $\mathcal{L}$ if the neural ODE started at pseudo-time s

with initial value $(z(s), \varphi(s) + \delta) = (z(s), \theta + \delta)$, where δ is an infinitesimal perturbation. Since the loss $\mathcal{L}$ ultimately only depends on $z(1)$, we have

$$\frac{\partial \mathcal{L}}{\partial \varphi(1)} = 0.$$

The gradient we wish to obtain is

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \varphi(0)}.$$

By the same reasoning as before, we have

$$\begin{aligned}\dot{a}_{\text{aug}}(s) &= -a_{\text{aug}}(s)\frac{\partial f_{\text{aug}}}{\partial z_{\text{aug}}}(z_{\text{aug}}(s), s) \\ &= -\begin{bmatrix} a(s) & b(s) \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial z}(z(s), \varphi(s), s) & \frac{\partial f}{\partial \theta}(z(s), \varphi(s), s) \\ 0 & 0 \end{bmatrix}.\end{aligned}$$

Multiplying out this leads to the stated result. □

Finally, we are ready to describe the algorithm to perform backpropagation with the neural ODE.

Step 1. With initial condition $z(0)$, call an ODE solver to compute $z(1)$.

Step 2. With initial condition $(z(1), a(1), b(1))$, with $z(1)$ computed from step 1, $a(1) = \frac{\partial \mathcal{L}}{\partial z(1)}$, and $b(1) = 0$, call ODE solver (backwards in pseudo-time) to compute $(z(0), a(0), b(0))$. Return $b(0) = \frac{\partial \mathcal{L}}{\partial \theta}$.

The ODE solver call of Step 2 requires the values of $\{z(s)\}_{s\in[0,1]}$ at appropriate discrete points. One option is to store the values of $\{z(s)\}_{s\in[0,1]}$ computed in Step 1. Another option, is to compute $\{z(s)\}_{s\in[0,1]}$ anew from $z(1)$ together with the computation of $\{a(s)\}_{s\in[0,1]}$ and $\{b(s)\}_{s\in[0,1]}$. This second option, described in the above algorithm, is much more memory (storage) efficient, although it does require slightly more computation.

Bibliography

[1] Lee K. Jones. A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *The Annals of Statistics*, 20(1):608–613, 1992.

[2] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[3] Gilles Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire Analyse fonctionnelle (dit*, pages 1–12, 1981.

[4] Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International conference on machine learning*, pages 2979–2987. PMLR, 2017.

[5] Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.