An Overview of Gradient Descent in Reproducing Kernel Hilbert Spaces

Carson James
Texas A&M University

Abstract

This report will give an overview of various tools used to perform gradient descent in the context of reproducing kernel Hilbert spaces and will present an application to Gaussian processes.

1 Banach Spaces

1.1 Bounded Linear Maps

Definition 1.1.1. Let X be a normed vector space. Then X is said to be a **Banach space** if X is complete.

Definition 1.1.2. Let X, Y be a normed vector spaces and $T : X \to Y$ a linear map. Then T is said to be **bounded** if there exists $C \geq 0$ such that for each $x \in X$,

$$\|Tx\| \leq C\|x\|$$

We define

$$L(X, Y) = \{T : X \to Y : T \text{ is linear and bounded}\}$$

Definition 1.1.3. Let $X_1, \ldots, X_n$ and Y be a normed vector spaces and $T : \prod_{j=1}^{n} X_j \to Y$ a multilinear linear map. Then T is said to be **bounded** if there exists $C \geq 0$ such that for each $(x_j)_{j=1}^n \in \prod_{j=1}^{n} X_j$,

$$\|T(x_1, \ldots, x_n)\| \leq C\|x_1\| \ldots \|x_n\|$$

We define $L^n(X_1, \ldots, X_n; Y)$ to be the set of $T : X \to Y$ such that T is multilinear and bounded. If $X_1, \ldots, X_n = X$, we write $L^n(X, Y)$ in place of $L^n(X, \ldots, X; Y)$.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Remark 1.1.4. Let X and Y be normed vector spaces. We may identify $L(X, L(X, \ldots, L(X, Y)) \ldots)$ and $L^n(X, Y)$ via the isometric isomorphism given by $\phi \mapsto \psi_\phi$ where

$$\psi_\phi(x_1, x_2, \ldots, x_n) = \phi(x_1)(x_2) \ldots (x_n)$$

Definition 1.1.5. Let X be a normed vector space over $\mathbb{R}$. We define the **dual space of** X, denoted X^*, by $X^* = L(X, \mathbb{R})$. Let $T : X \to \mathbb{R}$. Then T is said to be a **bounded linear functional on** X if $T \in X^*$.

1.2 Differentiation

Definition 1.2.1. Let X, Y be a banach spaces, $A \subset X$ open, $f : A \to Y$ and $x_0 \in A$. Then f is said to be **(1-st order) Frechet differentiable at** x_0 if there exists $Df(x_0) \in L(X, Y)$ such that,

$$f(x_0 + h) = f(x_0) + Df(x_0)(h) + o(\|h\|) \qquad \text{as } h \to 0$$

If f is Frechet differentiable at x_0, we define the **Frechet derivative of** f **at** x_0 to be $Df(x_0)$. We say that f is **(1-st order) Frechet differentiable** if for each $x_0 \in A$, f is Frechet differentiable at x_0.
If f is Frechet differentiable, we define the **Frechet derivative** of f, denoted $Df : A \to L(X, Y)$, by

$$x \mapsto Df(x)$$

Continuing inductively, if f is $(n-1)$-th order Frechet differentiable, f is said to be n-th order Frechet differentiable at x_0 if $D^{n-1}f$ is Frechet differentiable at x_0. We define $D^n f(x_0) = D(D^{n-1}f)(x_0)$.

Remark 1.2.2. Note that $D^n f(x_0) \in L^n(X, Y)$.

1.3 Calculus

Remark 1.3.1. The tools used to obtain the following results:

- Frechet Derivative
- Bochner Integral

- Hahn-Banach Theorem

Result 1.3.1. *Let X, Y be Banach spaces and $f \in L(X,Y)$. Then f is Frechet differentiable and for each $x_0 \in X$, $Df(x_0) = f$.*

Result 1.3.2. *Let X, Y, Z be Banach spaces, $f : X \to Y$, $g : Y \to Z$ and $x_0 \in X$. If f is Frechet differentiable at x_0 and g is Frechet differentiable at $f(x_0)$, then $g \circ f$ is Frechet differentiable at x_0 and*

$$D(g \circ f)(x_0) = Dg(f(x_0)) \circ Df(x_0)$$

Result 1.3.3. ***Mean Val Thm:***
Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f : A \to Y$. If f is Frechet differentiable, then for each $x, y \in A$, there exists $t \in (0,1)$ such that

$$\|f(x) - f(y)\| \leq \|Df(tx + (1-t)y)\|\|x - y\|$$

Result 1.3.4. *Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f : A \to Y$. Suppose that f is Frechet differentiable. If for each $x \in A$, $Df(x) = 0$, then f is constant.*

Result 1.3.5. *Let X, Y be a Banach spaces, $A \subset X$ open and convex and $f, g : A \to Y$. Suppose that f and g are Frechet differentiable. If $Df = Dg$, then there exists $c \in Y$ such that $f = g + c$.*

Result 1.3.6. *Let X be a Banach spaces, $A \subset X$ open, $f : A \to \mathbb{R}$ and $x_0 \in A$. Suppose that f is Frechet differentiable at x_0. If f has a local minimum at x_0, then $Df(x_0) = 0$.*

Result 1.3.7. ***Fundy Thm of Calc:***
Let Y be a separable Banach space and $f \in C^1_Y(a,b)$. Then for each $x, x_0 \in (a,b)$, $x_0 < x$ implies that

1. *f' is Bochner integrable on $(x_0, x]$*
2. $$f(x) - f(x_0) = \int_{(x_0,x]} f' dm$$

Result 1.3.8. ***Taylor's Theorem:***
Let Y be a separable Banach space, $A \subset X$ open and convex, $f \in C^n_Y(A)$ and $x_0 \in A$. Then

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{1}{k!} D^k f(x_0)(h, \ldots, h) + o(\|h\|^n)$$

as $h \to 0$

2 Hilbert Spaces

2.1 Introduction

Definition 2.1.1. Let H be an inner product space. Then H is said to be a **Hilbert space** if H is complete with respect to the norm induced by the inner product.

Remark 2.1.2. We will be assuming the Hilbert space is real.

Result 2.1.1. ***Cauchy Schwarz Ineq:***
Let H be an inner product space. Then for each $x, y \in H$, $|\langle x, y\rangle| \leq \|x\|\|y\|$ with equality iff $x \in \text{span}(y)$.

2.2 Riesz Representation Theorem

Definition 2.2.1. Let H be a Hilbert space. Define $\phi : H \to H^*$ by $x \mapsto x^*$ where

$$x^* y = \langle x, y\rangle$$

Result 2.2.1. *Let H be a Hilbert space. Then $\phi : H \to H^*$ defined above is an isometric isomorphism.*

Definition 2.2.2. Let H be a Hilbert space, $f : H \to \mathbb{R}$ and $x_0 \in H$. Suppose that f is Frechet differentiable at x_0 so that $Df(x_0) \in H^*$. We define the **gradient of f at x_0**, denoted $\nabla f(x_0) \in H$, by

$$\nabla f(x_0) = \phi^{-1} Df(x_0)$$

That is, $\nabla f(x_0)$ is the unique element of H such that for each $y \in H$,

$$\langle \nabla f(x_0), y\rangle = Df(x_0)(y)$$

Result 2.2.2. *Let H be a Hilbert space, $f : H \to \mathbb{R}$ and $x_0 \in H$. If f is Frechet differentiable at x_0, then*

$$\arg\min_{\|h\| \leq 1} Df(x_0)(h) = -\|\nabla f(x_0)\|^{-1} \nabla f(x_0)$$

Remark 2.2.3. In the context of Hilbert spaces, the gradient allows us generalize the gradient descent method for minimization.

The idea is as follows. If $f : H \to \mathbb{R}$ is Frechet differentiable. Then

$$f(x_0 + h) \approx f(x_0) + \langle \nabla f(x_0), h\rangle$$

for h near 0. Taking $h = -\eta \nabla f(x_0)$ for some small $\eta > 0$ insures that h is close to 0 and h is in the direction of steepest descent of $Df(x_0)(v)$ which causes $f(x_0 + h) < f(x_0)$.

3 Convex Analysis

3.1 Results

Result 3.1.1. *Let X be a vector space, $A \subset X$ convex, $f : A \to \mathbb{R}$ convex and $x_0 \in A$. Then f has a local minimum at x_0 iff f has a global minimum at x_0.*

Result 3.1.2. *Let X be a vector space, $A \subset X$ convex and $f : A \to \mathbb{R}$ strictly convex. If f has a local minimum, then there exists a unique $x_0 \in A$ such that $f(x_0) = \min_{x \in A} f(x)$.*

Result 3.1.3. *Let X be a Banach space, $A \subset X$ open and convex, $f : A \to \mathbb{R}$ convex, $x_0 \in A$. Suppose that f is 2nd order Frechet differentiable. If for each $x_0 \in A$, $D^2 f(x_0) \in L^2(X, \mathbb{R})$ is positive semi definite (resp. pos. def.), then f is convex (resp. strictly convex).*

Remark 3.1.1. By positive definite, we mean $D^2 f(x_0)(h, h) > 0$ for $h \neq 0$.

4 Reproducing Kernel Hilbert Spaces

4.1 RKHS's

Definition 4.1.1. Let T be a set and $H \subset \mathbb{R}^T$ a hilbert space. For $t \in T$, we define the **evauluation functional at** t, denoted $L_t : H \to \mathbb{R}$, by

$$L_t(f) = f(t)$$

The space H is said to be a **reproducing kernel Hilbert space (RKHS)** if for each $t \in T$, $L_t \in H^*$ (i.e. L_t is bounded).
If H is an RKHS, the Riesz representation theorem implies that for each $t \in T$, there exists $K_t \in H$ such that for each $f \in H$, $\langle K_t, f \rangle = f(t)$.
If H is an RKHS, we define the **reproducing kernel** associated to H, denoted $K_H : T^2 \to \mathbb{R}$, by

$$K_H(s, t) = \langle K_s, K_t \rangle$$

Result 4.1.1. *Let T be a set and $K : T^2 \to \mathbb{R}$. If K is symmetric and positive definite, then there exists a unique reproducing kernel Hilbert space $H \subset \mathbb{R}^T$ such that $K_H = K$.*

Result 4.1.2. ***Representer Theorem:***
Let T be a set, $K : T^2 \to \mathbb{R}$ a symmetric, postivie definite kernel on T, $H \subset \mathbb{R}^T$ the corresponding RKHS, $t = (t_j)_{j=1}^n \subset T$ and $y = (y_j)_{j=1}^n \subset \mathbb{R}$.
Define $L : H \to \mathbb{R}$ by

$$L(f) = \sum_{j=1}^{n} (y_j - f(t_j))^2 + \lambda \|f\|^2$$

Put $\hat{f} = \arg\min_{f \in H} L(f)$.
Then there exist $(\hat{\alpha}_j)_{j=1}^n \subset \mathbb{R}$ such that

$$\hat{f}(t) = \sum_{j=1}^{n} \hat{\alpha}_j K(t, t_j)$$

Remark 4.1.2. Define $A \in \mathbb{R}^{n \times n}$ by $A_{i,j} = K(t_i, t_j)$. Some regular calculus shows that $\hat{\alpha} = (A + \lambda I)^{-1} y$

Remark 4.1.3. If $(A + \lambda I)^{-1}$ is too expensive to compute, we may try other methods to find a minimum of L. The next section explores one such method, namely gradient descent.

5 Gradient Descent

5.1 Update Derivation

Remark 5.1.1. Using the setup from the previous section, define $Q : H \to \mathbb{R}$ by

$$Q(f) = \sum_{j=1}^{n} (y_j - f(t_j))^2$$

We can write rewrite $Q(f)$ as

$$Q(f) = \|L_t(f) - y\|_2^2$$

where $L_t \in L(H, \mathbb{R}^n)$ is given by

$$L_t(f) = (f(t_j))_{j=1}^n$$

Writing out $Q(f_0 + h)$, we see that

$$\begin{aligned} Q(f_0 + h) &= \|L_t(f_0) - y\|_2^2 + 2(L_t(f_0) - y)^T L_t(h) \\ &\quad + \|L_t(h)\|_2^2 \\ &= Q(f_0) + \text{linear funct of } h \\ &\quad + \text{bilinear funct of } (h, h) \end{aligned}$$

Equating terms from Taylors theorem, we see that $D^2 Q(f_0)(h, h) = 2\|L_t(h)\|_2^2$, which is p.s.d. So Q is convex. Since norms are convex and $\lambda \geq 0$, L is convex.

Remark 5.1.2. Similar to before, writing out $L(f_0 + h)$, we get

$$\begin{aligned} L(f_0 + h) &= L(f_0) + 2(L_t(f_0) - y)^T L_t(h) + 2\lambda \langle f_0, h \rangle \\ &\quad + o(\|h\|^2) \end{aligned}$$

So

$$\begin{aligned} DL(f_0)(h) &= 2(L_t(f_0) - y)^T L_t(h) + 2\lambda \langle f_0, h \rangle \\ &= 2 \sum_{j=1}^{n} (f_0(t_j) - y_j) \langle K_{t_j}, h \rangle + 2\lambda \langle f_0, h \rangle \\ &= \left\langle 2 \left[\sum_{j=1}^{n} (f_0(t_j) - y_j) K_{t_j} + \lambda f_0 \right], h \right\rangle \end{aligned}$$

Hence

$$\nabla L(f_0) = 2 \left[\sum_{j=1}^{n} (f_0(t_j) - y_j) K_{t_j} + \lambda f_0 \right]$$

Therefore the gradient descent update reads as follows:

$$\begin{aligned} f_{t+1} &= f_t - \eta \nabla L(f_t) \\ &= (1 - 2\eta\lambda) f_t - 2\eta \left[\sum_{j=1}^{n} (f_0(t_j) - y_j) K_{t_j} \right] \end{aligned}$$

6 Applications to Gaussian Processes

6.1 Introduction

Remark 6.1.1. Let T be a set and $x = (x_j)_{j=1}^n \in T^n$, $y = (y_j)_{j=1}^n \in \mathbb{R}^n$. Recall that if

$$\begin{aligned} y_i &= f(x_i) + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \\ f &\sim GP(0, c) \end{aligned}$$

Then

$$f|x, y \sim GP(\tilde{\mu}, \tilde{c})$$

where

$$\tilde{\mu}(t) = c(t, x)[c(x, x) + \sigma^2 I]^{-1} y$$

and

$$\tilde{c}(s, t) = c(s, t) - c(s, x)[c(x, x) + \sigma^2 I]^{-1} c(x, t)$$

6.2 Posterior Mean and Covariance

Remark 6.2.1. If $(c(x, x) + \sigma^2 I)^{-1}$ is too expensive to compute, we may set up the following convex optimization problems to approximate the posterior mean and posterior covariance functions via our gradient descent algorithm:

- $$\tilde{\mu}(t) = \arg\min_{f \in H} \sum_{j=1}^{n} (y_j - f(t_j))^2 + \sigma^2 \|h\|_H$$

- Fixing $t \in T$,

$$\hat{c}(\cdot, t) = \arg\min_{f \in H} \sum_{j=1}^{n} (c(t_j, t) - f(t_j))^2 + \sigma^2 \|h\|_H$$

where H is the RKHS corresponding to the p.d. kernel c.
The first optimization problem lets us approximate $\tilde{\mu}$ directly by gradient descent and the second optimization problem lets us approximate $\tilde{c}(t)$ by finding $\hat{c}(\cdot, t)$ via gradient descent and the computing $\tilde{c}(s, t) = c(s, t) - \hat{c}(s, t)$.

Acknowledgements

Remark 6.2.2. Essentially all the proofs are in the notes on my Github page listed in the references or on Wikipedia. I got the material about Gaussian processes from Dr. Pati's STAT 605 class.

References

1. Analysis Notes
2. Integration Notes
3. RKHS Wikipedia Page
4. Representer Theorem Wikipedia Page