Learning in Hilbert spaces : a new framework

January 27, 2022

Abstract

1 Intuition

Let us start with a simple example : a regression problem. Say you are modelling a relation between two quantities X and Y by

$$Y = h(X) + E \ , \tag{1}$$

where E is some noise. You have access to n observations of (X, Y) in the form $(X_i, Y_i)_{i=1,..,n}$ and you want to learn the function h, or at least approximate it. How do you proceed? *A priori*, you don't know the form of the objective function h, so you will try some non-parametric regression algorithms (nearest neighbors, regression trees, kernel regression, etc). But if you knew that h belonged to some known parametric family, say $(h_\beta)_{\beta\in\mathbb{R}^d}$, you will simply perform a gradient descent. You will associate a risk function to the model, for example the mean squared error

$$R(\beta) = \mathbb{E}(h_\beta(X) - Y)^2 \ . \tag{2}$$

and your gradient update will have the form

$$\beta^{t+1} = \beta^t - \frac{2\alpha}{n}\sum_{i=1}^{n}(h_{\beta^t}(X_i) - Y_i)\partial_\beta h_{\beta^t}(X_i) \ , \ \forall t = 1, .., T \ . \tag{3}$$

where α is the step-size and T is the number of iterations. A convexity analysis would then assure you that your algorithm converges to the true β.

Can we perform a gradient descent even in the non-parametric case?

Well, if we define a similar risk

$$R(h) = \mathbb{E}(h(X) - Y)^2 =: \mathbb{E}L(h, X, Y) \ , \tag{4}$$

then we know that the Fréchet gradient, when it exists, is a natural extension of the classical gradient in finite dimension. If h lives in some space H. For a fixed x, y, the Fréchet gradient $\nabla L(h, x, y)$ of $L(\cdot, x, y)$ in some point h is a bounded linear form on H and should satisfy

$$\lim_{g\to 0}\frac{|L(h+g, x, y) - L(h, x, y) - \nabla L(h, x, y)(g)|}{\|g\|_H} = 0 \ . \tag{5}$$

Moreover, if H is an Hilbert space, we know by Riesz theorem that $\nabla L(h,x,y)$ has a representative (which we also call $\nabla L(h,x,y)$) in H and we can write

$$\lim_{g\to 0} \frac{|L(h+g)-L(h)-\langle \nabla L(h,x,y), g\rangle_H|}{\|g\|_H} = 0 \ . \tag{6}$$

If this representative was known, then we can easily perform a gradient update in H

$$h^{t+1} = h^t - \frac{2\alpha}{n}\sum_{i=1}^n \nabla L(h^t, X_i, Y_i) \ .$$

Do we know what the representative of the Fréchet gradient is ?

One important point to keep in mind is that this representative depends on the space H. When performing finite-dimensional gradient descent, we often seek to have bounds on the L_2 norm of $\beta - \beta^*$, if β^* is the true parameter. So assume that H is for example $L^2(0,1)$, and for a fixed $x \in [0,1]$, consider the functional $F : H \to \mathbb{R}$ defined by

$$F(h) = h(x) \ . \tag{7}$$

If we know how to compute the Fréchet gradient of this functional, then we can easily compute the Fréchet gradient of L via a composition argument. Let $g \in H$, since F is linear, we have

$$\frac{F(h+g)-F(h)}{\|g\|_H} = \frac{g(x)}{\|g\|_H} \ .$$

So we expect the Fréchet gradient of F in h to be

$$\nabla F(h)(g) = g(x) \ .$$

The linear form $g \to g(x)$ doesn't have a representative in L^2 (it is not bounded for example). So we are stuck here and can not compute a Fréchet gradient. One solution is to change the space H. Let us consider the Sobolev space

$$H = \{g, \text{ s.t } g(t) = \int_0^1 g'(s)\mathrm{d}s \text{ with } g' \in L^2(0,1)\} \ . \tag{8}$$

This space is nice for two reasons. First we can embody it with the scalar product

$$\langle g, h\rangle_H = \int_0^1 h'(s)g'(s)\mathrm{d}s \ , \tag{9}$$

and we know that

$$\forall x \in [0,1] \ \forall g \in H, |g(x)| \leq \|g\|_H \ . \tag{10}$$

So this means that one, the linear form $g \to g(x)$ becomes continuous in this space, and two, we have

$$g(x) = \int_0^1 \mathbf{1}_{[0,x]}(s)g'(s)\mathrm{d}s = \langle \int_0^{\cdot} \mathbf{1}_{[0,x]}(s)\mathrm{d}s, g\rangle_H \ . \tag{11}$$

So, in this space, we expect the representative of the Fréchet gradient of F to be

$$\nabla F(h)(\cdot) = \int_0^{\cdot} \mathbf{1}_{[0,x]}(s)\mathrm{d}s = \cdot \wedge x \ . \tag{12}$$

It happens to be that this gradient we found coincides with the Malliavin Gradient, a tool developed for analysis of stochastic variations.

What is the Malliavin Gradient?

In the space H, the Malliavin gradient is first defined for stochastic integrals. Consider the canonical Brownian motion

$$\begin{aligned} B:\ \Omega = \mathcal{C}_0(0,1) \to & \mathcal{C}_0(0,1) \\ w \to & B(w) = \Big(t \to w(t)\Big)\ , \end{aligned}$$

and the stochastic integrals, defined for all $g \in H$ as a random variable $\delta g = \int_0^1 g'(s)\mathrm{d}B_s$. Notice that δg can be seen as a random variable or as a functional defined from $\mathcal{C}_0$ to $\mathbb{R}$.

Then if f is a smooth function from $\mathbb{R}^m \to \mathbb{R}$ for some integer $m \geq 1$ and $h_1, .., h_m$ are vectors in H, we define the Malliavin gradient of the functional

$$\begin{aligned} F:\ \mathcal{C}_0(0,1) \to & \mathbb{R} \\ w \to & F(w) = f(\delta h_1(w), ..., \delta h_m(w))\ , \end{aligned} \tag{13}$$

as the H valued random variable

$$\begin{aligned} DF:\ \mathcal{C}_0(0,1) \to & H \\ w \to & DF(w) = \sum_{j=1}^{m} \partial_j f(\delta h_1(w), ..., \delta h_m(w)) h_j\ . \end{aligned} \tag{14}$$

Basically, we are saying that the gradient of δg is g, and we extend the notion to smooth functions of δg. The Malliavin gradient can then be extended by a completeness argument via the norm

$$\|F\|_{1,2}^2 = \mathbb{E}F^2 + \mathbb{E}\|DF\|_H^2\ . \tag{15}$$

With this new tool in hand, it is easy to see that

$$h(x) = \Big(\int_0^1 \mathbf{1}_{[0,x]}(s)\mathrm{d}B_s\Big)(h) = \delta\Big(\cdot \wedge x\Big)(h)\ . \tag{16}$$

Indeed, the second term is equal to $B_x(h) = h(x)$. Therefore, we have

$$D(h \to h(x)) = \cdot \wedge x\ , \tag{17}$$

which coincides with the Fréchet gradient found above. In fact, in the space H, the Malliavin and Fréchet gradient are the same for all functionals of the form (13) (i.e $DF = \nabla F$). And if we go back to our squared loss for the regression problem, we have

$$DL(h,x,y) = \nabla L(h,x,y) = 2(h(x) - y)(\cdot \wedge x)\ . \tag{18}$$

So, we can perform a gradient descent in H via the update

$$h^{t+1}(\cdot) = h^t(\cdot) - \frac{2\alpha}{n}\sum_{i=1}^{n}(h^t(X_i) - Y_i)(\cdot \wedge X_i)\ . \tag{19}$$

Malliavin Gradient or Fréchet Gradient?

To highlight the differences between the two, we give an example of two functions, the first one is Fréchet differentiable but doesn't have a Malliavin Gradient. The second one has a Malliavin Gradient but isn't Fréchet differentiable. Consider

$$\begin{aligned} E : H &\to \mathbb{R} \\ g &\to e^{g(1)^2} \end{aligned}$$

Then E is clearly Fréchet differentiable, as it is a composition of the exponential and the evaluation functional at 1. But F does not have a Malliavin gradient. This is because of the integrability imposed in (24). Because of the Gaussian nature of the Brownian motion, we don't have $\mathbb{E}(e^{B_1^2}) < \infty$. Consider now

$$\begin{aligned} E : \mathcal{C}_0(0,1) &\to \mathbb{R} \\ g &\to |g(1)| \ . \end{aligned}$$

This function is not Fréchet differentiable, because it is not Fréchet differentiable in 0, but it has a Malliavin Gradient

$$DE(g)(\cdot) = sign(g(1)) \ (\cdot \wedge 1) \ .$$

Again, this is because of the norm used in (24). Because of the L_2 integrability, the Malliavin Gradient does not care what happens in an event that is negligible, such as $\{g(1) = 0\}$.

Why is the Malliavin Gradient interesting?

There are three reasons why. We were lucky to find the Fréchet representative in the case of a regression problem with a squared loss, but generally speaking, the Fréchet gradient is only defined as a linear form and one can not always find the representative in H. The Malliavin Gradient, on the other hand, is defined directly as an element of H. The second reason is more important. In Malliavin Calculus, there exists an integration by parts formulae. Let $F : \mathcal{C}_0(0,1) \to \mathbb{R}$ such that F has a Malliavin Gradient DF. Then for all $u \in H$, we have

$$\mathbb{E}(\langle DF, h\rangle_H) = \mathbb{E}(F\delta h) \ . \tag{20}$$

To highlight that DF is a random variable in H, let us write the equation above as

$$\mathbb{E}(\langle DF(w), u\rangle_H) = \mathbb{E}(F(w) \ \delta u(w)) \ .$$

Letting $u(\cdot) = \cdot \wedge z$. We have then that for all $z \in [0,1]$

$$\mathbb{E}(DF(w)(z)) = \mathbb{E}(F(w) \ (B_z(w))) \ .$$

This basically means that we can approximate the Malliavin gradient of F, only using F. Let us incorporate this in our learning setting. For all $x, y \in [0,1]$ and $\gamma > 0$, let

$$\tilde{L}(h,x,y) = \mathbb{E}L(h + \gamma B, x, y) \ .$$

Then, by the integration by parts property, we have that for all $x \in [0,1]$

$$D\tilde{L}(h)(z) = \frac{1}{\gamma}\mathbb{E}(L(h + \gamma B, x, y)B_z) \ .$$

A gradient-free update?

Since the *new* loss $\tilde{L}$ is close to L when γ is small. We can now perform a gradient-free gradient descent. Say we have access to n trajectories of the Brownian motion in the form $(B^i)_{i=1,...,n}$, then the gradient-free update takes the form

$$h^{t+1}(\cdot) = h^t(\cdot) - \frac{2\alpha}{n} \sum_{i=1}^{n} L(h^t + \gamma B^i, X_i, Y_i) B^i_\cdot \ . \tag{21}$$

This update holds for all losses L that have a Malliavin Gradient, so all losses that fall into the set of functions (13), or that are limits of such functions in the sense (24). Finally, Malliavin Calculus is a general theory, defined for general Hilbert spaces that benefit from certain structures. The properties used above would still be true and our approach would have the potential to work in other Hilbert spaces, multi-dimensional ones for example, which would be interesting from a learning perspective.

2 Regression problem in higher dimension

Let us consider the same problem as before, but this time, we allow X to be a d-dimensional vector in the space $[0,1]^d$. Define now the space H as

$$H = \{g, g(t) = \int_0^{t_1} \int_0^{t_d} g'(u_1, .., u_d) \mathrm{d}u_1 ... \mathrm{d}u_d, \text{ with } g' \in L^2([0,1]^d)\} \ .$$

Notice that as before, H has an RKHS structure with the kernels $(\Phi_t)_{t\in[0,1]}$ defined by

$$\Phi_t(\cdot) = \Pi_{i=1}^d (\cdot_i \wedge t_i) \ .$$

What does the Malliavin Gradient look like in higher dimension?

Consider the canonical Brownian sheet

$$\begin{aligned} B: \ \Omega = \mathcal{C}_0([0,1]^d) &\to \mathcal{C}_0([0,1]^d) \\ w &\to B(w) = (t \to w(t)) \ , \end{aligned}$$

which is the only centered Gaussian process with continuous trajectories and covariance given by $\mathbb{E}(B_t B_s) = \Phi_t(s)$. The stochastic integral is then defined for all $h \in H$

$$\delta h = \int_{[0,1]^d} h'(t) \mathrm{d}B_t \ .$$

Again, the Malliavin Gradient of δh is then defined as h and then extended to functionals

$$\begin{aligned} F: \ \mathcal{C}_0([0,1]^d) &\to \mathbb{R} \\ w &\to F(w) = f(\delta h_1(w), ..., \delta h_m(w)) \ , \end{aligned} \tag{22}$$

as the H valued random variable

$$\begin{aligned} DF: \ \mathcal{C}_0([0,1]^d) &\to H \\ w &\to DF(w) = \sum_{j=1}^{m} \partial_j f(\delta h_1(w), ..., \delta h_m(w)) h_j \ . \end{aligned} \tag{23}$$

The Malliavin gradient can then be extended by a completeness argument via the norm

$$\|F\|_{1,2}^2 = \mathbb{E}F^2 + \mathbb{E}\|DF\|_H^2 \ . \tag{24}$$

With this in hand, it is easy to see that for all $x \in [0,1]^d$

$$D(h \to h(x)) = \Phi_x(\cdot) \ . \tag{25}$$

It follows that for the squared loss, we have for all $x, y \in [0,1]^d \times [0,1]$

$$DL(h,x,y)(\cdot) = 2(h(x) - y)\Phi_x(\cdot) \ . \tag{26}$$

The interested reader can check that the above gradient is again the same as the Fréchet gradient in the space H.

A gradient update in higher dimension?

In the space H, we can therefore perform the gradient update

$$h^{t+1}(\cdot) = h^t(\cdot) - \frac{2\alpha}{n}\sum_{i=1}^{n}(h^t(X_i) - Y_i)(\Phi_{X_i}(\cdot)) \ . \tag{27}$$

Moreover, using the integration by parts property, we can also perform the gradient-free update

$$h^{t+1}(\cdot) = h^t(\cdot) - \frac{2\alpha}{n}\sum_{i=1}^{n} L(h^t + \gamma B^i, X_i, Y_i)B^i_\cdot \ . \tag{28}$$

which again, holds for general losses and not only the squared loss.

Is the space H too restrictive?

Let g be continuous function defined on some bounded domain of $\mathbb{R}^d$ and let $\varepsilon > 0$. Then, via translation and shrinking techniques, g can be seen as a function on $[\varepsilon, 1-\varepsilon]^d$. Moreover, g can be continuously extended to $[0,1]^d$ such that $g(t) = 0$ if one of the coordinates of t vanishes. Since the space H is dense in the space of continuous functions that vanish whenever one of the coordinates vanishes. Our gradient descent will learn a function $\tilde{g}$ that is uniformly close to g. For instance, in our regression problem, we can learn all continuous functions as long as the support of the distribution of X is finite.

3 Gradient descent for general models and loss functions

3.1 Learning setting

3.2 Learning algorithms

3.3 Analysis

References