Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Functional Gradient Descent through Directional Derivatives

Caio Lins[1], Daniel Csillag[2], Yuri F. Saporito[3], Bernardo F. P. da Costa[4]
FGV EMAp, RJ

Many problems in science and engineering can be understood as a risk minimization procedure over a suitable linear space, such as regression tasks in statistics, or finding solutions to partial differential equations and inverse problems. See [1] for other examples. The natural spaces where the solutions to these problems live are often infinite dimensional, which leads to tractability problems. In response, many solution approaches involve, in one way or another, a reformulation within a parametric, finite-dimensional setting. In boundary value problems for PDEs, for example, finite element methods use a weak formulation of the PDE over a suitable discretization of the domain to arrive at a finite-dimensional linear system. Even more modern ideas, such as the Physics Informed Neural Networks (PINNs) put forward in [2], involve representing the solution to the PDE through parametric function in the form of a neural network, although the number of parameters can be quite high.

In this work, we wish to directly tackle the problem of risk minimization in an infinite dimensional space, and our strategy will be to perform **(stochastic) gradient descent** within this space. The main difficulty comes from the fact that, in our problems of interest, it is not possible to compute the (stochastic) gradient exactly, so we must employ approximation strategies.

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space. The problem we want to solve is

$$\min_{h \in \mathcal{H}} \mathcal{R}(h), \tag{1}$$

where $\mathcal{R}$ is a **risk functional** $\mathcal{R} : \mathcal{H} \to \mathbb{R}$ which we assume to be Fréchet differentiable with its derivative being denoted by[5] $\mathcal{DR} : \mathcal{H} \to \mathcal{H}^*$. Note that, for $h \in \mathcal{H}$, $\mathcal{DR}(h) : \mathcal{H} \to \mathbb{R}$ is a continuous linear functional on $\mathcal{H}$ and, by the Riesz Representation Theorem, admits a representation as the inner product with a member of $\mathcal{H}$, which we call the **gradient** of $\mathcal{R}$ at h and denote by $\nabla\mathcal{R}(h)$.

Our intent is to tackle problem (1) through gradient descent, however, as mentioned before, we are not able to exactly compute (stochastic) gradients. The object which we **can** compute is the **directional derivative** of $\mathcal{R}$ at $h \in \mathcal{H}$ in any direction $v \in \mathcal{H}$, which we denote by[6] $D\mathcal{R}(h)(v)$. Fortunately, this is enough to obtain useful gradient estimators:

Quadratic approximation In this method, we exploit the following fact[7]:

$$\nabla\mathcal{R}(h) = \operatorname*{arg\,min}_{\varphi \in \mathcal{H}} -\langle \nabla\mathcal{R}(h), \varphi \rangle + \frac{1}{2}\|\varphi\|^2 \tag{2}$$

[1] caio.peixoto@fgv.br

[2] daniel.csillag@fgv.br

[3] yuri.saporito@fgv.br

[4] bernardo.paulo@fgv.br

[5] We denote the dual space of a Banach space X by X^*.

[6] Contrast this notation with the **calligraphic** $\mathcal{D}$ used for the Fréchet derivative.

[7] This characterization of the gradient is the one used by Mirror Descent methods when applied in Hilbert spaces with a Bregman divergence given by the standard metric in $\mathcal{H}$.

Notice that the expression on the RHS only depends on $\nabla\mathcal{R}(h)$ through its inner product with φ, which is precisely the directional derivative at h in the direction φ. Hence, we can rewrite it as

$$\arg\min_{\varphi\in\mathcal{H}} -\langle\nabla\mathcal{R}(h),\varphi\rangle + \frac{1}{2}\|\varphi\|^2 = \arg\min_{\varphi\in\mathcal{H}} -D\mathcal{R}(h)(\varphi) + \frac{1}{2}\|\varphi\|^2 \tag{3}$$

Since we are not able to conduct this minimization exactly over $\mathcal{H}$, we obtain an estimator for $\nabla\mathcal{R}(h)$ by minimizing over a subclass $\mathcal{F}\subset\mathcal{H}$. Our gradient estimator $\widehat{\nabla\mathcal{R}(h)}_{\text{QA}}$ is then given by

$$\widehat{\nabla\mathcal{R}(h)}_{\text{QA}} = \arg\min_{\varphi\in\mathcal{F}\subset\mathcal{H}} -D\mathcal{R}(h)(\varphi) + \frac{1}{2}\|\varphi\|^2 \tag{4}$$

For example, if $\mathcal{H}$ is a space of differentiable functions we may take $\mathcal{F}$ to be the set of neural networks with a given architecture. □

Inner product matching In this method, obtain an approximation of $\nabla\mathcal{R}(h)$ by trying to match its average inner product with randomly selected elements of $\mathcal{H}$. Mathematically, let μ_h be a probability measure defined on[8] $(\mathcal{H},\mathcal{B}(\mathcal{H}))$. For example, one may think of μ_h as being a Gaussian measure. If the covariance operator of μ_h is injective, then

$$\nabla\mathcal{R}(h) = \arg\min_{\varphi\in\mathcal{H}} \mathbb{E}_{\psi\sim\mu_h}\left[(\langle\nabla\mathcal{R}(h),\psi\rangle - \langle\varphi,\psi\rangle)^2\right]. \tag{5}$$

Since this expectation is hard to compute exactly we employ a Monte Carlo estimate. Given samples i.i.d. $\psi_1,\ldots,\psi_N \overset{\text{iid}}{\sim} \mu_h$, the estimator $\widehat{\nabla\mathcal{R}(h)}_{\text{IPM}}$ is given by

$$\widehat{\nabla\mathcal{R}(h)}_{\text{IPM}} = \arg\min_{\varphi\in\mathcal{F}\subset\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N}\left(D\mathcal{R}(h)(\psi_i) - \langle\varphi,\psi\rangle\right)^2. \qquad \square$$

Letting $\widehat{\nabla\mathcal{R}(h)}$ denote any of these estimators, we can then compute approximate solutions to (1) through

$$h_{m+1} = h_m - \alpha_m\widehat{\nabla\mathcal{R}(h_m)}, \tag{6}$$

where (α_m) is a sequence of learning rates. In our work, we apply this methodology to boundary value problems in PDEs.

References

[1] Y. R. Fonseca and Y. F. Saporito. **Statistical Learning and Inverse Problems: A Stochastic Gradient Approach**. 2022. arXiv: 2209.14967 [stat.ML].

[2] M. Raissi, P. Perdikaris, and G. Karniadakis. **Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations**. 2017. arXiv: 1711.10561 [cs.AI].

[8] We denote the Borel σ-algebra in a topological space X by $\mathcal{B}(X)$.