From PDE-Constrained Optimization to DNNs

Ali Hamza Abidi, Syed & Andreas Mang
Department of Mathematics, University of Houston, Houston, TX, USA

Teaser: Our goal was the design and analysis of effective numerical schemes for training deep neuronal networks (**DNN**s) based on optimal control formulations.

In the present work we explore numerical methods inspired by optimal control theory to train image classifiers [1]. In a first step, we consider a prototypical formulation of a variational optimization problem governed by an elliptic dynamical system [2]. We will discuss the numerical treatment and study some of the mathematical operators. Subsequently, we present the optimal control formulation for training DNNs and derive some expressions for the associated optimality conditions. In our future work, we plan to extend these optimality conditions and device a numerical scheme for the DNN training problem, similar to the scheme developed for the prototypical problem.

Optimal Control Problems

In the present work we study numerical methods for solving problems of the general form [3]

$$\underset{u\in\mathcal{U},\ w\in\mathcal{W}}{\text{minimize}}\ \ \text{dist}(u_{\text{pred}}, y^\delta) + \text{reg}(w) \quad \text{subject to } \mathcal{C}(w,u) = 0.$$

Here, $u \in \mathcal{U}$ is the state variable, $w \in \mathcal{W}$ is the control variable, and $c : \mathcal{W}\times\mathcal{U} \to \mathcal{Q}$, $\mathcal{C}(w,u) := \mathcal{A}(w)u - q$ is a dynamical system, i.e., a partial differential equation (PDE). The functional dist measures the discrepancy between the predicted state $u_{\text{pred}} := \mathcal{A}^{-1}(w)q$ and the observable (data) y^δ, with *parameter-to-obervation map* $\mathcal{F}(w) := \mathcal{A}^{-1}(w)q$. The functional reg is a Tikhonov regularization model introduced to alleviate issues with the ill-posedness of the optimization problem stated above [4].

Prototypical Example

We consider the following PDE-constrained optimization problem [2]

$$\begin{aligned}\underset{u\in\mathcal{U},\ w\in\mathcal{W}}{\text{minimize}}\ \ & \frac{1}{2}\int_\Omega (u-y^\delta)^2\mathrm{d}x + \frac{\alpha}{2}\|w\|_{\mathcal{W}}^2\\ \text{subject to}\ \ & \mathcal{C}(w,u) = \nabla\cdot w\nabla u - q = 0\end{aligned}$$

subject to Dirichlet boundary conditions $u = 0$, with gradient operator ∇, divergence operator $\nabla\cdot$, and regularization parameter $\alpha > 0$. We consider $q(x) = 1$ if $x_1 \in (0.4, 0.6), x_2 \in (0.4.0.6)$ and $q(x) = 0$ otherwise, the parameter function $w_{\text{true}}(x) = 2.2 + 2\sin(\pi x_1)\sin(\pi x_2)$, and $y^\delta = \mathcal{A}(w_{\text{true}})^{-1}q$ (see Figure 1).

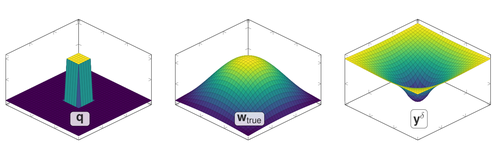

Figure 1: *Visualization of the problem data. From left to right: The right hand side q, the parameter function w_{true}, and the solution of the forward problem for this input data y^δ.*

We discretize the functions u, w, and q on a cell-centered grid of size $n_1 \times n_2$. We consider a finite-volume discretization and arrange all variables in lexicographical ordering to obtain vectors $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{q} \in \mathbb{R}^n$, with $n = n_1 n_2$. The discrete optimization problem is given by

$$\underset{\mathbf{u}\in\mathbb{R}^n,\mathbf{w}\in\mathbb{R}^n}{\text{minimize}}\ \ \frac{h}{2}\|\mathbf{u}-\mathbf{y}^\delta\|_2^2 + \frac{\alpha h}{2}\|\mathbf{L}\mathbf{w}\|_2^2 \ \text{ subject to } \ \mathbf{c}(\mathbf{u},\mathbf{w}) = \mathbf{0}$$

Here, $\mathbf{L} \in \mathbb{R}^{n\times n}$ denotes the regularization operator and $\mathbf{c}(\mathbf{u},\mathbf{w}) = \text{DSG}\,\mathbf{u} - \mathbf{q}$, $\text{DSG} = \text{DIV}\,\mathbf{S}\,\text{GRAD} \in \mathbb{R}^{n,n}$, where

$$\text{GRAD} = \begin{bmatrix} \boldsymbol{I}_2 \otimes \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \otimes \boldsymbol{I}_1 \end{bmatrix} \in \mathbb{R}^{m,n}$$

is the discretized gradient operator with $m = (n_1+1)n_2 + n_1(n_2+1)$. The matrix $\boldsymbol{S} = \text{diag}(\boldsymbol{s}) \in \mathbb{R}^{m,m}, \boldsymbol{s} = \boldsymbol{e}_m \oslash (\boldsymbol{A}(\boldsymbol{e}_n \oslash \boldsymbol{w})) \in \mathbb{R}^m$, $\boldsymbol{e}_n = \begin{bmatrix}1 \ \dots \ 1\end{bmatrix}^\mathsf{T} \in \mathbb{R}^n$, $\oslash : \mathbb{R}^n \to \mathbb{R}^n$, represents the discretization of $\boldsymbol{w}$ via harmonic averaging, and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_2 \otimes \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \otimes \boldsymbol{I}_1 \end{bmatrix} \in \mathbb{R}^{m,n}$$

with $\boldsymbol{A}_i \in \mathbb{R}^{n_i+1,n_i}$, $\boldsymbol{I}_i = \text{diag}(1\ \dots\ 1) \in \mathbb{R}^{n_i,n_i}$, $i = 1, 2$. The matrix

$$\text{DIV} = \begin{bmatrix}\boldsymbol{I}_2 \otimes \boldsymbol{D}_1 & \boldsymbol{D}_2 \otimes \boldsymbol{I}_1\end{bmatrix} \mathbb{R}^{n,m} \tag{1}$$

represents the divergence operator with $\boldsymbol{I}_i$ as defined above and $\boldsymbol{D}_i \in R^{n_i+1,n_i}$ is the one-dimensional derivative operator along the coordinate direction x_i, $i = 1, 2$. The sensitivity $\boldsymbol{J}(\boldsymbol{w})$ and its adjoint are given by

$$\boldsymbol{J} := \boldsymbol{J}(\boldsymbol{w}) = -\boldsymbol{Q}\boldsymbol{C}_u^{-1}\boldsymbol{C}_w \quad \text{and} \quad \boldsymbol{J}^\mathsf{T} := \boldsymbol{J}(\boldsymbol{w})^\mathsf{T} = -\boldsymbol{C}_w^\mathsf{T}\boldsymbol{C}_u^{-\mathsf{T}}\boldsymbol{Q}^\mathsf{T},$$

respectively, with derivative matrices $\boldsymbol{C}_w := \mathrm{d}_w\boldsymbol{c}(\boldsymbol{u},\boldsymbol{w}) \in \mathbb{R}^{n,n}$ and $\boldsymbol{C}_u := \mathrm{d}_u\boldsymbol{c}(\boldsymbol{u},\boldsymbol{m}) \in \mathbb{R}^{n,n}$ (see Figures 2 & 3). Given the PDE constraint $\boldsymbol{c}(\boldsymbol{u},\boldsymbol{w}) = 0$ and a candidate model $\boldsymbol{w}$, we can solve $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{w})$. With this, we can eliminate the constraint from the optimization problem and reformulate the equality constrained problem as an unconstrained problem of the form

$$\underset{\boldsymbol{u}\in\mathbb{R}^n}{\text{minimize}}\ \ \left\{ f(\boldsymbol{w}) := \frac{h}{2}\|\boldsymbol{u}(\boldsymbol{w}) - \boldsymbol{y}^\delta\|_2^2 + \frac{\alpha h}{2}\|\boldsymbol{L}\boldsymbol{w}\|_2^2 \right\}.$$

A necessary condition for a minimizer of this problem is given by the non-linear system $\mathrm{d}_w f(\boldsymbol{w}^\star) = h\boldsymbol{J}^\mathsf{T}(\boldsymbol{u}(\boldsymbol{w}^\star) - \boldsymbol{y}^\delta) + \alpha h\boldsymbol{L}^\mathsf{T}\boldsymbol{L}\boldsymbol{w}^\star = 0$.

Numerical Optimization

We have developed a matrix-free Newton–Krylov method for its solution. We use an *iterative line search scheme* of the form

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \mu_k\boldsymbol{B}_k\mathrm{d}_w f(\boldsymbol{w}_k), \quad k = 1, 2, \dots$$

Here, $k \in \mathbb{N}$ is the iteration index and $\mu_k \in (0, 1]$ is determined using a backtracking line search [5]. The search direction is given by $\boldsymbol{s}_k := -\boldsymbol{B}_k\mathrm{d}_w f(\boldsymbol{w}_k)$. We consider *Newton's method* with $\boldsymbol{B}_k = \boldsymbol{H}^{-1}$, $\boldsymbol{H} := \mathrm{d}_{ww} f(\boldsymbol{w}) = h\boldsymbol{J}(\boldsymbol{w})^\mathsf{T}\boldsymbol{J}(\boldsymbol{w}) + \mathrm{d}_w(\boldsymbol{J}(\boldsymbol{w})^\mathsf{T}\boldsymbol{r}) + \alpha h\boldsymbol{L}^\mathsf{T}\boldsymbol{L}$, where $\boldsymbol{r} := \boldsymbol{u}(\boldsymbol{w}) - \boldsymbol{y}^\delta$ is constant. We invert the Hessian matrix using a matrix-free, conjugate gradient method with a superlinear forcing sequence. As a stopping criterion, we consider the relative reduction of the norm of the gradient.

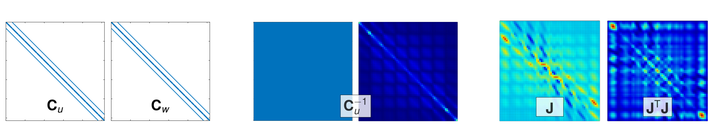

Figure 2: *Left: Sparsity pattern of the Jacobians $\boldsymbol{C}_u$ (1210 non-zero entries) and $\boldsymbol{C}_u$ (1220 non-zero entries) for $d = 2$, $n_i = 16$. Middle: Sparsity pattern (for $d = 2$, $n_i = 16$) and visualization of the entries (for $d = 2$, $n_i = 8$) of the inverse of $\boldsymbol{C}_u$ (65 500 non-zero entries; the matrix is dense). Right: Visualization of the entries of the sensitivities $\boldsymbol{J}$ and $\boldsymbol{J}^\mathsf{T}\boldsymbol{J}$ for $d = 2$, $n_i = 8$.*

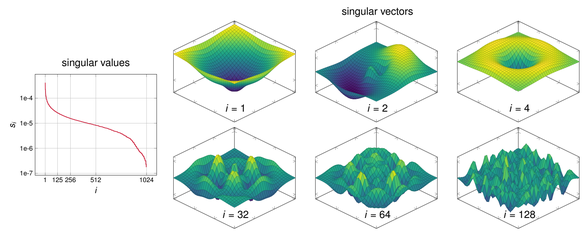

Figure 3: *Singular value decomposition $\boldsymbol{J} = \boldsymbol{U}\Sigma\boldsymbol{V}^\mathsf{T}$ of the sensitivity matrix $\boldsymbol{J}$ of the size 1020×1020. Left: Singular values of the operator $\boldsymbol{J}$. Right: Singular vectors corresponding to the singular values s_i of $\Sigma = \text{diag}(s_i)$ for $i \in \{1, 2, 4, 32, 64, 128\}$.*

Optimal Control Formulation for DNNs

The **optimal control formulation** for training a deep neural network is given by [1]

$$\begin{aligned}\underset{\boldsymbol{\Phi}}{\text{minimize}}\ \ & \text{dist}(\boldsymbol{C}_{\text{pred}}, \boldsymbol{C}) + \alpha\,\text{reg}(\boldsymbol{W}, \boldsymbol{\mu}, \{\boldsymbol{K}_i\}_{i=1}^n, \{b_i\}_{i=1}^n)\\ \text{subject to}\ \ & \boldsymbol{Y}_{j+1} = \boldsymbol{Y}_j + h\sigma(\boldsymbol{Y}_j\boldsymbol{K}_j + \boldsymbol{b}_j),\end{aligned} \tag{2}$$

$j = 0, 1, \cdots, n-1$. Here, $\text{dist} : \mathbb{R}^{s,m}\times\mathbb{R}^{s,m} \to \mathbb{R}$ measures the discrepancy between the predicted classification $\boldsymbol{C}_{\text{pred}} \in \mathbb{R}^{s,m}$ and the labels (data) $\boldsymbol{C} \in \{0, 1\}^{s,m}$. The unknowns $\boldsymbol{\Phi}$ of the optimization problem (2) are the weights $\boldsymbol{K}_i \in \mathbb{R}^{n,n}$ and biases b_i of the "ResNET" forward propagation and the weights $\boldsymbol{W} \in \mathbb{R}^{n,m}$ and biases $\boldsymbol{\mu} \in \mathbb{R}^m$ that parameterize the classifier. Consequently, $\boldsymbol{\Phi} := \{\boldsymbol{W}, \boldsymbol{\mu}, \{\boldsymbol{K}_i\}_{i=1}^n, \{b_i\}_{i=1}^n\}$. The prediction $\boldsymbol{C}_{\text{pred}}$ is computed according to $\boldsymbol{C}_{\text{pred}} = g(\boldsymbol{Y}_n\boldsymbol{W} + \boldsymbol{e}_s\otimes\mu)$, where $\boldsymbol{Y}_n$ is the final state computed by solving the forward propagation, $\boldsymbol{e}_s := (1, \dots, 1)^\mathsf{T} \in \mathbb{R}^s$, and $g : \mathbb{R}^{s,m} \to \mathbb{R}^{s,m}$ is the so called hypothesis function. The partial derivatives of the loss function $\|g(\boldsymbol{Y}_n\boldsymbol{W} + \boldsymbol{e}_s\otimes\mu) - \boldsymbol{C}\|_F^2/s$ are given by $\mathrm{d}_W\|g(\boldsymbol{Y}_n\boldsymbol{W}) - \boldsymbol{C}\|_F^2 = 2\boldsymbol{Y}_n^\mathsf{T}(g'(\boldsymbol{Y}_n\boldsymbol{W})\odot(g(\boldsymbol{Y}_n\boldsymbol{W}) - \boldsymbol{C}))/s$ and $\mathrm{d}_Y\|g(\boldsymbol{Y}_n\boldsymbol{W}) - \boldsymbol{C}\|_F^2 = 2(g'(\boldsymbol{Y}_n\boldsymbol{W})\odot(g(\boldsymbol{Y}_n\boldsymbol{W}) - \boldsymbol{C}))\boldsymbol{W}^\mathsf{T}/s$. If we consider the regularization operators $\text{reg}(\boldsymbol{K}) = \|\boldsymbol{K}\|_F^2/2 = \text{tr}(\boldsymbol{K}\boldsymbol{K}^\mathsf{T})/h$ and $\text{reg}(\boldsymbol{b}) = \|\boldsymbol{b}\|_2^2/h$ we obtain the derivatives $\mathrm{d}_K = \boldsymbol{K}$ and $\mathrm{d}_b = \boldsymbol{b}$, respectively. The derivatives of the dynamical system (derivation of the sensitivities) is more complicated and forms the basis of our current work.

Conclusions

We have explored the implementation of an optimization framework for optimal control problems governed by PDEs. We have developed a matrix-free, *Newton–Krylov method* globalized by an *Armijo linesearch* for numerical optimization. We have started to derive optimality conditions for the optimal control problem for training DNNs described in [1]. The derivation of the sensitivities for this problem forms the basis of our current work.

References

1. E. Haber & L. Ruthotto, Stable Architectures for Deep Neural Networks, Inverse Problems 34 014004, 2017.
2. E. Haber, Computational Methods in Geophysical Electromagnetics, SIAM, 2015.
3. F. Tröltzsch, Optimal Control of Partial Differential Equations: Theory, Methods, and Applications, American Mathematical Society, 2010.
4. H. W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Springer, 2000.
5. J. Nocedal & S. Wright, Numerical Optimization, Springer Science, 1999.