Differential Machine Learning – Appendix 3
Differential Regression

Brian Huge
brian.huge@danskebank.dk

Antoine Savine
antoine.savine@danskebank.dk

May 5, 2020

Introduction

Differential machine learning is discussed in detail in the working paper, in the context of deep learning, where its *unreasonable effectiveness* is illustrated with examples picked in real world applications and textbooks, like a basket option in a correlated Bachelier model in section 3.1.

In this appendix, we apply the same ideas in the context of classic regression on a fixed set of basis functions, and demonstrate equally remarkable results, illustrated with the same Bachelier basket example, with pricing and risk functions approximated by polynomial regression. Recall that the example from the paper is reproduced on a public notebook `https://github.com/differential-machine-learning/notebooks/blob/master/DifferentialML.ipynb`. We posted another notebook *DifferentialRegression.ipynb* with the regression example, where the formulas of this document for standard, ridge and differential regression are implemented and compared.

Like standard and ridge regression, differential regression is performed in closed form and lends itself to SVD stabilization. Unlike ridge regression, differential regression provides strong regularization *without bias*. It follows that there is no bias-variance tradeoff with differential regression, in particular, the sensitivity to regularization strength is virtually null. As illustrated in the notebook, differential regression vastly outperforms Tikhonov regularization, even when the Tikhonov parameter is optimized by cross validation at the cost of additional data consumption. Differential regression doesn't consume additional data besides a training set augmented with differentials as explained in the paper. It doesn't necessitate additional regularization or hyperparameter optimization by cross validation.

The exercise is to perform a classic least square linear regression $\widehat{Y} = \mu_Y + (\phi - \mu_\phi)\,\beta$, where the columns of $\phi = \phi(X)$ are basis functions (e.g. monomials, excluding constant) of known inputs X (also excluding constant, with examples in rows and inputs in columns), given a column vector Y of the corresponding targets, where μ_Y is the mean of Y and the row vector μ_ϕ contains the means of the columns of ϕ. To simplify notations, we denote $\phi \equiv \phi - \mu_\phi$ and $Y \equiv Y - \mu_Y$. Classic least squares finds β by minimization of the least square errors:

$$\beta = \arg\min_\beta \|Y - \phi\beta\|^2$$

The analytic solution, also called 'normal equation':

$$\beta = \left(\phi^T\phi\right)^{-1}\phi^T Y$$

is known to bear unstable results, the matrix $\phi^T\phi$ often being near singular (certainly so with monomials of high degree of correlated inputs). This is usually resolved with SVD regression. We prefer the (very similar) eigenvalue regression, which we recall first, and then, extend to ridge (Tikhonov) regularization and finally

differential regression. Parts 1 and 2 are summaries of classic results. Part 3 is new. After β is learned, the value approximation for an input row vector x is given by $\widehat{y} = \phi(x)\beta$ and the derivative approximations are given by:

$$\widehat{y}_j = \phi_j(x)\beta$$

where subscripts denote partial derivatives to input number j.

1 SVD regression, eigenvalue variant

Perform the eigenvalue decomposition $\phi^T\phi = PDP^T$.

Denote $D^{-\frac{1}{2}}$ the diagonal matrix whole diagonal elements are the elements of the diagonal matrix D, raised to power -0.5 , when they exceed a threshold (say, 10^{-8} times the mean trace of D), and zero otherwise.

Denote $\widetilde{\phi} = \phi P D^{-\frac{1}{2}}$ and perform the least square minimization in the orthonormal basis:

$$\widetilde{\beta} = \arg\min_{\widetilde{\beta}} \left\| Y - \widetilde{\phi}\widetilde{\beta} \right\|^2$$

The normal equation is stable in the orthonormal basis:

$$\widetilde{\beta} = \left(\widetilde{\phi}^T\widetilde{\phi}\right)^{-1}\widetilde{\phi}^T Y$$

It is easy to see that $\widetilde{\phi}^T\widetilde{\phi}$ is a diagonal matrix with diagonal elements 1 corresponding to significant eigenvalues, and 0 corresponding to insignificant ones. With the convention $\left(\widetilde{\phi}^T\widetilde{\phi}\right)^{-1} = \widetilde{\phi}^T\widetilde{\phi}$ (invert the significant diagonal elements and zero the insignificant ones), we get:

$$\widetilde{\beta} = D^{-\frac{1}{2}}P^T\phi^T Y$$

(notice, $D^{-\frac{1}{2}}$ zeroes the lines corresponding to insignificant eigenvalues so there is no need to left multiply by $\widetilde{\phi}^T\widetilde{\phi}$.) Hence: $\widehat{Y} = \widetilde{\phi}\widetilde{\beta} = \phi P D^{-1} P^T\phi^T Y = \beta Y$ where $D^{-1} = \left(D^{-\frac{1}{2}}\right)^2$ has diagonal elements inverse of the significant eigenvalues in D, zero for the insignificant eigenvalues, and:

$$\boxed{\beta = PD^{-1}P^T\phi^T Y}$$

2 Tikhonov (ridge) regularization

Classic regression works best with regularization, the most common classic form of which is ridge regression, also called Tikhonov regularization, which adds a penalty on the norm of β in the objective cost.

$$
\begin{aligned}
\beta &= \arg\min_{\beta}\left[\|Y-\phi\beta\|^2+\lambda^2\|\beta\|^2\right] \\
&= \arg\min_{\beta}\left[\left\|Y-\left(\phi PD^{-\frac{1}{2}}\right)\left(D^{\frac{1}{2}}P^T\beta\right)\right\|^2+\lambda^2\|\beta\|^2\right] \\
&= PD^{-\frac{1}{2}}\arg\min_{\gamma}\left[\left\|Y-\widetilde{\phi}\gamma\right\|^2+\lambda^2\left\|PD^{-\frac{1}{2}}\gamma\right\|^2\right] \\
&= PD^{-\frac{1}{2}}\arg\min_{\gamma}\left[\left\|Y-\widetilde{\phi}\gamma\right\|^2+\lambda^2\left\|D^{-\frac{1}{2}}\gamma\right\|^2\right] \\
&= PD^{-\frac{1}{2}}\left[\widetilde{\phi}^T\widetilde{\phi}+\lambda^2D^{-1}\right]^{-1}\widetilde{\phi}^TY \\
&= PD^{-\frac{1}{2}}\left[\widetilde{\phi}^T\widetilde{\phi}+\lambda^2D^{-1}\right]^{-1}D^{-\frac{1}{2}}P^T\phi^TY \\
&= P\Lambda^{-1}P^T\phi^TY
\end{aligned}
$$

where Λ^{-1} has diagonal elements $\frac{1}{D_{ii}+\lambda^2}$ where D_{jj} is significant, zero otherwise. And we get:

$$\boxed{\beta(\lambda)=P\Lambda(\lambda)^{-1}P^T\phi^TY}$$

The Tikhonov parameter λ can be found e.g. by cross validation:

$$\lambda=\arg\min_{\lambda}\|Y_V-\phi_V\beta(\lambda)\|^2$$

where $\phi_V=\phi(X_V)$, (X_V,Y_V) form a validation set of independent examples and $\beta(\lambda)$ is the result of a ridge regression over the training set with Tikhonov parameter λ, obtained with the boxed formula above. The objective function can be expanded:

$$
\begin{aligned}
f(\lambda) &= \|Y_V-\phi_V\beta(\lambda)\|^2 \\
&= (Y_V-\phi_V\beta(\lambda))^T(Y_V-\phi_V\beta(\lambda)) \\
&= \left(Y_V-\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY\right)^T\left(Y_V-\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY\right) \\
&= \left({Y_V}^T-Y^T\phi P\Lambda(\lambda)^{-1}P^T{\phi_V}^T\right)\left(Y_V-\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY\right) \\
&= {Y_V}^TY_V-Y^T\phi P\Lambda(\lambda)^{-1}P^T{\phi_V}^TY_V-{Y_V}^T\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY \\
&\quad +Y^T\phi P\Lambda(\lambda)^{-1}P^T{\phi_V}^T\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY
\end{aligned}
$$

Since ${Y_V}^TY_V$ doesn't depend on λ, we minimize:

$$
\begin{aligned}
g(\lambda) &= Y^T\phi P\Lambda(\lambda)^{-1}P^T{\phi_V}^T\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY \\
&- Y^T\phi P\Lambda(\lambda)^{-1}P^T{\phi_V}^TY_V \\
&- {Y_V}^T\phi_VP\Lambda(\lambda)^{-1}P^T\phi^TY \\
&= K^T\Lambda(\lambda)^{-1}M\Lambda(\lambda)^{-1}K-K^T(\lambda)^{-1}L-L^T\Lambda(\lambda)^{-1}K \\
&= K^T\Lambda(\lambda)^{-1}M\Lambda(\lambda)^{-1}K-2K^T(\lambda)^{-1}L \\
&= K^T\Lambda(\lambda)^{-1}\left[M\Lambda(\lambda)^{-1}K-2L\right]
\end{aligned}
$$

where

$$K=P^T\phi^TY\,\{n\times1\},L=P^T{\phi_V}^TY_V\,\{n\times1\}\text{ and }M=P^T{\phi_V}^T\phi_VP\,\{n\times n\}$$

Optimization may be efficiently performed by a classic one-dimensional minimization procedure.

3 Differential Regression

In addition to inputs X and labels Y, we have derivatives labels Z whose columns Z_j are the differentials of Y to X_j. Denote ϕ_j the matrix of derivatives of the basis functions ϕ wrt X_j. Linear regression makes value predictions $\widehat{Y} = \phi\beta$ and derivatives predictions $\widehat{Z}_j = \phi_j\beta$. We now minimize a cost combining value and derivatives errors:

$$\beta = \arg\min_{\beta} \left[\|Y - \phi\beta\|^2 + \sum_j \lambda_j \|Z_j - \phi_j\beta\|^2 \right]$$

where $\lambda_j = \lambda^2 \frac{\|Y\|^2}{\|Z_j\|^2}$ (norms are computed across examples) ensures that the components of the cost are of similar magnitudes. The hyperparameter λ has little effect and generally left to 1.

It is not hard to see that this minimization is analytically solved with the adjusted normal equation:

$$\beta = \left(\phi^T\phi + \sum_j \lambda_j {\phi_j}^T \phi_j \right)^{-1} \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right)$$

This is, again, a theoretical equation, unstable in practice. As before, we change basis by eigenvalue decomposition of:

$$\phi^T\phi + \sum_j \lambda_j {\phi_j}^T \phi_j = PDP^T$$

– beware, notations have changed so P and D denote different (respectively unitary and diagonal) matrices than before. Changing basis as before: $\widetilde{\phi} = \phi P D^{-\frac{1}{2}}$ (where, as previously, $D^{-\frac{1}{2}}$ has zero diagonal elements where the eigenvalues in D are insignificant) we notice that:

$$\widetilde{\phi}_j \equiv \frac{\partial \widetilde{\phi}}{\partial X_j} = \frac{\partial \left(\phi P D^{-\frac{1}{2}} \right)}{\partial X_j} = \frac{\partial \phi}{\partial X_j} P D^{-\frac{1}{2}} = \phi_j P D^{-\frac{1}{2}}$$

Performing the minimization in the 'tilde' basis:

$$\widetilde{\beta} = \arg\min_{\widetilde{\beta}} \left[\left\| Y - \widetilde{\phi}\widetilde{\beta} \right\|^2 + \sum_j \lambda_j \left\| Z_j - \widetilde{\phi}_j \widetilde{\beta} \right\|^2 \right]$$

We have the normal equation:

$$\begin{aligned}
\widetilde{\beta} &= \left(\widetilde{\phi}^T\widetilde{\phi} + \sum_j \lambda_j \widetilde{\phi}_j^T \widetilde{\phi}_j \right)^{-1} \left(\widetilde{\phi}^T Y + \sum_j \lambda_j \widetilde{\phi}_j^T Z_j \right) \\
&= \left[\left(D^{-\frac{1}{2}} P^T \right) \left(\phi^T\phi + \sum_j \lambda_j {\phi_j}^T \phi_j \right) \left(P D^{-\frac{1}{2}} \right) \right]^{-1} \left[\left(D^{-\frac{1}{2}} P^T \right) \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right) \right] \\
&= D^{-\frac{1}{2}} P^T \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right)
\end{aligned}$$

Predicted values are given by:

$$\widehat{Y} = \widetilde{\phi}\widetilde{\beta} = \phi P D^{-\frac{1}{2}} D^{-\frac{1}{2}} P^T \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right) = \phi P D^{-1} P^T \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right) = \phi\beta$$

where $D^{-1} = \left(D^{-\frac{1}{2}}\right)^2$ is defined as previously (with zeroes on insignificant eigenvalues) and:

$$\boxed{\beta = P D^{-1} P^T \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right)}$$

Note that this is all consistent, in particular, derivatives predictions are given by:

$$\widehat{Z}_j = \phi_j \beta = \phi_j P D^{-1} P^T \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right) = \left(\phi_j P D^{-\frac{1}{2}} \right) \left[D^{-\frac{1}{2}} P^T \left(\phi^T Y + \sum_j \lambda_j {\phi_j}^T Z_j \right) \right] = \widetilde{\phi}_j \widetilde{\beta}$$

Conclusion

We derived a normal equation (SVD style) for differential regression (in the sense of the working paper's differential machine learning) and verified its effectiveness in a public demonstration notebook. Differential regularization vastly outperforms classic variants, including ridge, and without consuming additional data or needing any form of additional regularization or cross validation. Just like Tikhonov regularization, differential regularization is analytic and extremely effective, as seen in the demonstration notebook. Unlike Tikhonov, differential regularization is unbiased, as demonstrated in another appendix.