Artificial Neural Networks

Lecture 6: Bias-Variance and the Need for Regularization

Dr Juan Neirotti

j.p.neirotti@aston.ac.uk

2023

Introduction

- *The solution to the regression problem is the estimation of the underlying generator of data.*
- $\mathcal{P}_{T,\boldsymbol{X}}(t,\boldsymbol{x})$, *where* $t \in \mathscr{T} \subset \mathbb{R}$, $\boldsymbol{x} \in \mathscr{X} \subset \mathbb{R}^d$.
- *By definition we have:*

$$\mathcal{P}_{\boldsymbol{T},\boldsymbol{X}}(\boldsymbol{t},x) = \mathcal{P}_{\boldsymbol{T}|\boldsymbol{X}}(\boldsymbol{T}=\boldsymbol{t}|\boldsymbol{X}=x)\mathcal{P}_{\boldsymbol{X}}(x)$$

$$\mathcal{P}_{\boldsymbol{X}}(x) = \int_{\mathscr{T}} \mathrm{d}\boldsymbol{t}\,\mathcal{P}_{\boldsymbol{T},\boldsymbol{X}}(\boldsymbol{t},x).$$

- *In order to make a prediction (on an output given an input) we need to model* $\mathcal{P}_{T|\boldsymbol{X}}(t|\boldsymbol{x})$.

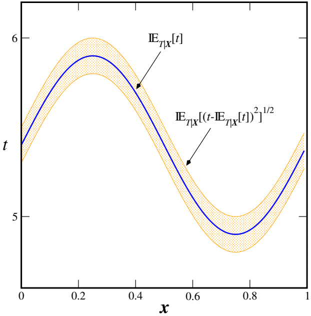


Figure: The full knowledge provided by $\mathcal{P}_{T,\boldsymbol{X}}(t,\boldsymbol{x})$ can be used to express the expected behavior of t in terms of $\boldsymbol{x}$.

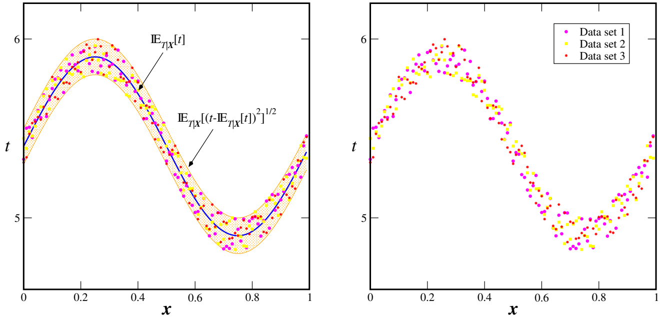


Figure: But we do not have $\mathcal{P}_{T,\boldsymbol{x}}(t,\boldsymbol{x})$. We have $\mathcal{D}_1$, $\mathcal{D}_2, \ldots$ that are samples drawn from $\mathcal{P}_{T,\boldsymbol{x}}(t,\boldsymbol{x})$.

Likelihood

- Several error measures are based on the *likelihood* $\mathcal{L}(\mathcal{D})$ of the data set $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1}^{N}$:

  $$\mathcal{L}(\mathcal{D}) = \prod_n \mathcal{P}_{T,\boldsymbol{x}}(t_n, \boldsymbol{x}_n)$$

  where we have assumed that the data points are drawn independently from the same distribution.
- Maximizing the likelihood is equivalent to minimizing the loss function defined as:

  $$L(\mathcal{D}) = -\log \mathcal{L} = -\sum_n \log \mathcal{P}_{T|\boldsymbol{x}}(t_n|\boldsymbol{x}_n) - \sum_n \log \mathcal{P}_{\boldsymbol{x}}(\boldsymbol{x}_n).$$

- The second term to the right hand side does not depend on the machine model being used, thus the effective loss becomes:

  $$L'(\mathcal{D}) = -\sum_n \log \mathcal{P}_{T|\boldsymbol{x}}(t_n|\boldsymbol{x}_n). \qquad (1)$$

Gaussian Noise

- Suppose the variable t is given by a combination of a deterministic process $h(\boldsymbol{x})$ plus a random variable ϵ drawn from a Gaussian distribution with zero mean and variance σ^2:

$$t = h(\boldsymbol{x}) + \epsilon$$

$$\mathcal{P}_E(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

- The deterministic function $h(\boldsymbol{x})$ is unknown, but is the only contribution to t that can be inferred from the data. Let as assume that there is an estimate $\varphi(\boldsymbol{x})$ that implements a model for $h(\boldsymbol{x})$. Such a model is associated with the following conditional probability of t :

$$\mathcal{P}_{T|\boldsymbol{X}}(t = \varphi(\boldsymbol{x})|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[t - \varphi(\boldsymbol{x})]^2}{2\sigma^2}\right\}. \tag{2}$$

Gaussian Noise and Maximum Likelihood

- By applying (1) with (2) we have that the log likelihood for a model with Gaussian Noise gives:

$$L'(\varphi) = \frac{1}{2\sigma^2}\sum_n \left[t_n - \varphi(\boldsymbol{x}_n)\right]^2 + \frac{N}{2}\ln(2\pi\sigma^2)$$

- The first term of the right hand side is the usual sum-of-squares error.
- Once optimized the model, by solving $\partial_{\sigma^2} L' = 0$, we can demonstrate that the variance satisfies:

$$\sigma^2 = \frac{1}{N}\sum_n \left[t_n - \varphi(\boldsymbol{x}_n)\right]^2 .$$

Noisy data

- We consider the cost function to be the sum of squares and that the size of the data set is *large*:

$$L(\varphi_{\mathcal{D}}) = \lim_{N\to\infty} \frac{1}{2N} \sum_{n=1}^{N} [\varphi(\boldsymbol{x}_n) - t_n]^2 \,,$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is the function implemented by the network.

- In such a limit we have that:

$$L(\varphi) = \frac{1}{2} \int \mathrm{d}t\, \mathrm{d}\boldsymbol{x}\, \mathcal{P}_{T|\boldsymbol{X}}(T = t|\boldsymbol{x}) \mathcal{P}(\boldsymbol{x}) \, [\varphi(\boldsymbol{x}) - t]^2$$

- Let us define the conditional averages:

$$\mathbb{E}_{T|\boldsymbol{X}}[t] := \int \mathrm{d}t\, \mathcal{P}_{T|\boldsymbol{X}}(t|\boldsymbol{x}) t, \qquad \mathbb{E}_{T|\boldsymbol{X}}[t^2] := \int \mathrm{d}t\, \mathcal{P}_{T|\boldsymbol{X}}(t|\boldsymbol{x}) t^2.$$

- Then

$$\begin{aligned}
L(\varphi) = & \frac{1}{2}\int \mathrm{d}t\,\mathrm{d}\boldsymbol{x}\,\mathcal{P}_{T|\boldsymbol{x}}(t|\boldsymbol{x})\mathcal{P}_{\boldsymbol{x}}(\boldsymbol{x})\left\{\left[\varphi(\boldsymbol{x}) - \mathbb{E}_{T|\boldsymbol{x}}[t]\right]^2\right. \\
& \left. + 2\left[\varphi(\boldsymbol{x}) - \mathbb{E}_{T|\boldsymbol{x}}[t]\right]\left[\mathbb{E}_{T|\boldsymbol{x}}[t] - t\right] + \left[\mathbb{E}_{T|\boldsymbol{x}}[t] - t\right]^2\right\} \\
= & \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\,\mathcal{P}(\boldsymbol{x})\left[\varphi(\boldsymbol{x}) - \mathbb{E}_{T|\boldsymbol{x}}[t]\right]^2 + \qquad (3) \\
& + \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\,\mathcal{P}(\boldsymbol{x})\left[\mathbb{E}_{T|\boldsymbol{x}}[t^2] - \mathbb{E}_{T|\boldsymbol{x}}[t]^2\right]. \qquad (4)
\end{aligned}$$

- Observe that the second contribution (4) is positive and does not depend on the model φ.
- The minimization of E is achieved for $\varphi_B(\cdot)$ such that $\varphi_B(\boldsymbol{x}) = \mathbb{E}_{T|\boldsymbol{x}}[t]$, known as the *Bayesian model.*

Finite data set

- Suppose that $|\mathcal{D}| = N < \infty$. In such a case, the quantity $\left[\varphi_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{T|\mathbf{x}}[t]\right]^2$ depends on the particular data set $\mathcal{D}$ used to train the model.
- We can eliminate this dependency by averaging over all possible data sets $\mathcal{D}$ with cardinality N. We denote such an average by $\mathbb{E}_{\mathcal{D}}[\cdot]$.
- Then:

$$\begin{aligned}
\left(\varphi_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{T|\mathbf{x}}[t]\right)^2 &= (\varphi_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\mathbf{x})] + \\
&+ \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\mathbf{x})] - \mathbb{E}_{T|\mathbf{x}}[t])^2 \\
&= (\varphi_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\mathbf{x})])^2 + \\
&+ \left(\mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\mathbf{x})] - \mathbb{E}_{T|\mathbf{x}}[t]\right)^2 \\
&+ 2\left(\varphi_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\mathbf{x})]\right) \times \\
&\times \left(\mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\mathbf{x})] - \mathbb{E}_{T|\mathbf{x}}[t]\right)
\end{aligned}$$

- By averaging both member over $\mathcal{D}$:

$$\mathbb{E}_{\mathcal{D}}\left[\left(\varphi_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}_{T|\boldsymbol{x}}[t]\right)^2\right] = \left(\mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})] - \varphi_B(\boldsymbol{x})\right)^2 + \tag{5}$$

$$+\, \mathbb{E}_{\mathcal{D}}\left[\left(\varphi_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})]\right)^2\right], \tag{6}$$

  where (5) is the squared *bias* term and (6) the *variance* term.
- The bias measures the extent to which the average over all data sets $\mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})]$ differs from the desired function $\mathbb{E}_{T|\boldsymbol{x}}[t] = \varphi_B(\boldsymbol{x})$.
- The variance measures the extent to which the network function $\varphi_{\mathcal{D}}(\boldsymbol{x})$ is sensitive to the particular choice of data set.
- Both contributions depend on $\boldsymbol{x}$.

Bias vs Variance

- We can eliminate the dependency over $\boldsymbol{x}$ by integrating:

$$
\begin{aligned}
(\mathrm{bias})^2 &= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x})\left(\mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})] - \varphi_B(\boldsymbol{x})\right)^2 \\
\mathrm{variance} &= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x})\mathbb{E}_{\mathcal{D}}\left[(\varphi_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})])^2\right].
\end{aligned}
$$

- Increasing the complexity of the model (number of parameters) reduces the bias but increase the sensibility of the model (variance) to the data set used (over fitting).

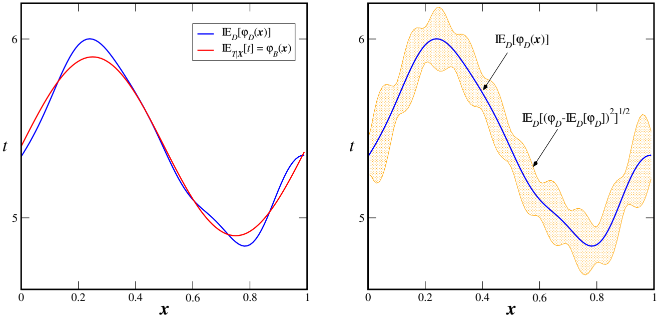


Figure: Bias and variance contributions respectively.

Error Decomposition for Regression

- The loss can be decomposed into the irreducible error, associated with the noise in the system, plus the bias squared plus the variance.

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[L(\varphi_{\mathcal{D}})] = {} & \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\, \mathcal{P}(\boldsymbol{x}) \left[\mathbb{E}_{T|\boldsymbol{x}}[t^2] - \mathbb{E}_{T|\boldsymbol{x}}[t]^2\right] + \\
& + \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\, \mathcal{P}(\boldsymbol{x}) \left(\mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})] - \varphi_B(\boldsymbol{x})\right)^2 + \\
& + \frac{1}{2}\int \mathrm{d}\boldsymbol{x} \mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x}) \mathbb{E}_{\mathcal{D}}\left[\left(\varphi_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}}[\varphi_{\mathcal{D}}(\boldsymbol{x})]\right)^2\right]
\end{aligned}
$$

Bias-Variance Like Decomposition for Classification

- Suppose that there is a process $\mathcal{P}_{\boldsymbol{X},T}(\boldsymbol{x},t)$ with $\boldsymbol{x} \in \mathscr{X} \subset \mathbb{R}^d$ and $t \in \{\pm 1\}$, and consider the functions $\mathscr{F} \ni \varphi : \mathscr{X} \rightarrow \{\pm 1\}$.
- Then the expectations satisfy:

$$\mathbb{E}_{\boldsymbol{X},T}\left[\Theta(-t\varphi(\boldsymbol{x}))\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{T|\boldsymbol{X}}\left[\Theta(-t\varphi(\boldsymbol{x}))\right]\right]$$

$$\begin{aligned}\mathbb{E}_{T|\boldsymbol{X}}\left[\Theta(-t\varphi(\boldsymbol{x}))\right] &= \sum_{t=-1,1} \mathcal{P}_{T|\boldsymbol{X}}(T=t|x)\Theta(-t\varphi(\boldsymbol{x})) \\ &= \mathcal{P}_{T|\boldsymbol{X}}(T \neq \varphi(\boldsymbol{x})|\boldsymbol{x})\end{aligned}$$

and the smallest possible generalization error occurs for the function:

$$\varphi_B(\boldsymbol{x}) = \operatorname*{argmin}_{\varphi\in\mathscr{F}} \mathbb{E}_{T|\boldsymbol{X}}\left[\Theta(-t\varphi(\boldsymbol{x}))\right].$$

.

- $\varphi_B(\cdot)$ is known as the Bayesian model, the model that better explains any set of data generated from the distribution $\mathcal{P}_{\boldsymbol{X},T}(\boldsymbol{x},t)$.
- It is straightforward from the definition of the minimum that:

  $$\varphi_B(\boldsymbol{x}) = \operatorname*{argmin}_{y\in\{\pm 1\}} \mathcal{P}_{T|\boldsymbol{X}}(T \neq y|\boldsymbol{x}),$$

  which is the minimum in the chances of making the incorrect classification y for a given condition $\boldsymbol{x}$.
- Equivalently:

  $$\varphi_B(\boldsymbol{x}) = \operatorname*{argmax}_{y\in\{\pm 1\}} \mathcal{P}_{T|\boldsymbol{X}}(T = y|\boldsymbol{x}),$$

  which is the maximum in the chances of making the correct classification y for a given condition $\boldsymbol{x}$.

- Observe that the Bayesian model $\varphi_B(\cdot)$ and the joint distribution $\mathcal{P}_{\boldsymbol{X},T}(\boldsymbol{x},t)$ are objects that can only be inferred from the knowledge brought forward by the data set.
- Now we introduce the sampling $\mathcal{D} = \{(\boldsymbol{x}_\ell, t_\ell)\}_{\ell=1}^{P} \in \{\mathscr{X} \times \{\pm 1\}\}^P$, from which we will infer the properties of $\mathcal{P}_{\boldsymbol{X},T}(\boldsymbol{x},t)$.
- If we consider the sampling $\mathcal{D}$ as a stochastic variable itself, we have that the model constructed from the data set $\varphi_{\mathcal{D}}(\boldsymbol{x})$ satisfies:

  $$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{T|\boldsymbol{X}}\left[\Theta(-t\varphi_{\mathcal{D}}(\boldsymbol{x}))\right]\right] = \mathcal{P}_{\mathcal{D},T|\boldsymbol{X}}(T \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}),$$

  i.e., the average generalization error (or the chances the model $\varphi_{\mathcal{D}}$ will produce the incorrect classification) over all possible data sets equals the conditional probability of $\varphi_{\mathcal{D}}(\boldsymbol{x})$ is incorrect.

- Thus:

$$
\begin{aligned}
M_{\mathcal{D}} &:= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{T|\boldsymbol{x}}\left[\Theta(-t\varphi_{\mathcal{D}}(\boldsymbol{x}))\right]\right] \\
&= 1 - \mathcal{P}_{\mathcal{D},T|\boldsymbol{x}}(T = \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}) \\
&= 1 - \mathcal{P}_{T|\boldsymbol{x}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x})\mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\varphi_B(\boldsymbol{x}) = \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}) - \\
&- \mathcal{P}_{T|\boldsymbol{x}}(T \neq \varphi_B(\boldsymbol{x})|\boldsymbol{x})\mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\varphi_B(\boldsymbol{x}) \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}) \\
&= 1 - \mathcal{P}_{T|\boldsymbol{x}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x})\left[1 - \mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\varphi_B(\boldsymbol{x}) \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x})\right] - \\
&- \mathcal{P}_{T|\boldsymbol{x}}(T \neq \varphi_B(\boldsymbol{x})|\boldsymbol{x})\mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\varphi_B(\boldsymbol{x}) \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}) \\
&= \mathcal{P}_{T|\boldsymbol{x}}(T \neq \varphi_B(\boldsymbol{x})|\boldsymbol{x}) + \\
&+ \mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\varphi_B(\boldsymbol{x}) \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x})\left[2\mathcal{P}_{T|\boldsymbol{x}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x}) - 1\right],
\end{aligned}
$$

which is the error of the best model plus the deviation of $\varphi_{\mathcal{D}}$ away from φ_B times the chances of φ_B being effectively correct.

.

- Suppose we construct an estimate $\hat{\mathcal{P}}_{\mathcal{D}}(T = y|\boldsymbol{x})$ for the class conditional probability, in such a way that:

$$\varphi_{\mathcal{D}}(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \{\pm 1\}} \hat{\mathcal{P}}_{\mathcal{D}}(T = y|\boldsymbol{x}).$$

- Therefore

$$\mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\varphi_B(\boldsymbol{x}) \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}) = \mathcal{P}_{\mathcal{D}|\boldsymbol{x}}(\hat{\mathcal{P}}_{\mathcal{D}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x}) < 0.5|\boldsymbol{x}),$$

which is the probability of the estimate $\hat{\mathcal{P}}_{\mathcal{D}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x})$ being less than fifty percent. Therefore the chances of the model $\varphi_{\mathcal{D}}$ to be different to φ_B are equal to the estimated chances of the model φ_B to produce the incorrect classification.

- *To illustrate the point let us assume that the estimate $\hat{\mathcal{P}}_{\mathcal{D}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x})$ is Normally distributed:*

$$
\begin{aligned}
\hat{\mathcal{P}}_{\mathcal{D}}(T = \varphi_{\boldsymbol{B}}(\boldsymbol{x})|\boldsymbol{x}) &= \hat{p} \\
\hat{p} &\sim \mathcal{N}\left(\hat{p}|\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right) \\
\mathcal{P}_{\mathcal{D}|\boldsymbol{X}}(\varphi_{\boldsymbol{B}}(\boldsymbol{x}) \neq \varphi_{\mathcal{D}}(\boldsymbol{x})|\boldsymbol{x}) &= \\
&= \int_0^{0.5} \mathrm{d}\hat{p}\mathcal{N}\left(\hat{p}|\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right) \\
&\approx \mathcal{H}\left(-\frac{0.5 - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)
\end{aligned}
$$

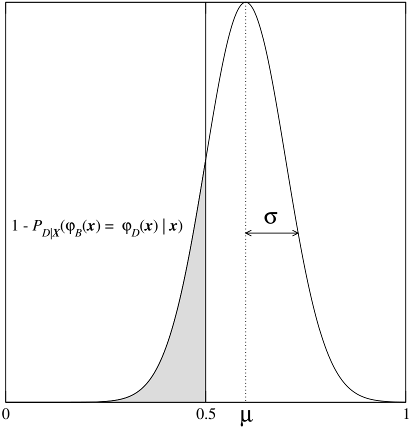


Figure: Probability distribution of $\hat{\mathcal{P}}_{\mathcal{D}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x})$.

.

The generalization error for classification becomes:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{T|\boldsymbol{X}}\left[\Theta(-t\varphi_{\mathcal{D}}(\boldsymbol{x}))\right]\right] = {} & \mathcal{P}_{T|\boldsymbol{X}}(T \neq \varphi_B(\boldsymbol{x})|\boldsymbol{x}) + \\ & + \mathcal{H}\left(-\frac{0.5-\mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)\times \\ & \times \left[2\mathcal{P}_{T|\boldsymbol{X}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{x}) - 1\right],\end{aligned}$$

when the expected probability estimate $\mathbb{E}_{\mathcal{D}}\left[\hat{\mathcal{P}}_{\mathcal{D}}(T = \varphi_B(\boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x})\right] = \mu(\boldsymbol{x})$ for the true majority class is greater than 0.5, reduction of the variance of the estimate ($\sigma^2(\boldsymbol{x})$) results in a decrease of the total misclassification error. If $\sigma \to 0$ then $\mathcal{H} \to 0$ and the expected generalization error gets reduced to the error of the optimal model. Conversely, when $\mu < 0.5$, a decrease in σ increases the total misclassification error. If $\sigma \to 0$,then $\mathcal{H} \to 1$ and the error is maximal.

.

Regularization: Weight Decay

- Let us consider the following error function:

  $$\tilde{E} = E + \nu\Omega(\boldsymbol{w})$$

  where the first term is the usual error measure, and the second term is the *penalty* term.
- The parameter $\nu \in \mathbb{R}^+$ controls the importance given to the penalty. The penalty term is related to the smoothness of the model (as a means to avoid over fitting).
- We aim to optimize a smooth model that produces a small error. The process of smoothing a model through an error penalty term is known as *regularization.*

Weight decay:

$$\Omega(\boldsymbol{w}) = \frac{1}{2}\langle \boldsymbol{w} \,|\boldsymbol{w}\,\rangle$$

where $\boldsymbol{w} \in \mathscr{W}$ are the parameters of the network.

.

- For a quadratic error we can write:

$$E(\boldsymbol{w}) = E_0 + \langle \boldsymbol{b} \,|\boldsymbol{w}\,\rangle + \frac{1}{2}\,\langle \boldsymbol{w}|\,\boldsymbol{H}\,|\boldsymbol{w}\rangle$$
$$\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \frac{\nu}{2}\,\langle \boldsymbol{w}\,|\boldsymbol{w}\,\rangle$$

  where the upper-index T indicates transpose, $\boldsymbol{b}$ is a constant vector and $\boldsymbol{H}$ is the Hessian (also constant) matrix.

The minimum of these expressions are obtained by solving:

$$|\boldsymbol{b}\rangle + \boldsymbol{H}\,|\boldsymbol{w}^{\star}\rangle = |\boldsymbol{0}\rangle$$
$$|\boldsymbol{b}\rangle + (\boldsymbol{H} + \nu\boldsymbol{I})\,|\tilde{\boldsymbol{w}}\rangle = |\boldsymbol{0}\rangle\,,$$

where $\boldsymbol{I}$ is the identity matrix of same dimensions as $\boldsymbol{H}$.

.

- Let us solve the Eigenvalue problem:

  $$\boldsymbol{H} \left|\lambda_k\right\rangle = \lambda_k \left|\lambda_k\right\rangle$$

  where $\left|\lambda_k\right\rangle$ are the (unit length) eigenvectors of $\boldsymbol{H}$ to the eigenvalue λ_k.
- We have then that, by using that $\sum_k \left|\lambda_k\right\rangle \left\langle\lambda_k\right| = \boldsymbol{I}$:

  $$\left|\boldsymbol{w}^\star\right\rangle = \sum_k \left|\lambda_k\right\rangle \left\langle\lambda_k \middle| \boldsymbol{w}^\star\right\rangle, \quad \tilde{\boldsymbol{w}} = \sum_k \left|\lambda_k\right\rangle \left\langle\lambda_k \middle| \tilde{\boldsymbol{w}}\right\rangle.$$
- Then, by the orthogonality of $\{\left|\lambda_k\right\rangle\}$:

  $$\sum_k \left[(\lambda_k + \nu) \left\langle\lambda_k \middle| \tilde{\boldsymbol{w}}\right\rangle - \lambda_k \left\langle\lambda_k \middle| \boldsymbol{w}^\star\right\rangle\right] \left|\lambda_k\right\rangle = \left|\boldsymbol{0}\right\rangle$$

  $$(\lambda_k + \nu) \left\langle\lambda_k \middle| \tilde{\boldsymbol{w}}\right\rangle - \lambda_k \left\langle\lambda_k \middle| \boldsymbol{w}^\star\right\rangle = 0$$

This implies that

$$\left\langle\lambda_k \middle| \tilde{\boldsymbol{w}}\right\rangle = \frac{\lambda_k}{\nu + \lambda_k} \left\langle\lambda_k \middle| \boldsymbol{w}^\star\right\rangle. \tag{7}$$

.

- From (7) we see that for the larger eigenvalues $\lambda_k \gg \nu$, which represent the more relevant components of the error landscape, the regularize model and the normal model produce similar results ($\langle \lambda_k | \tilde{\boldsymbol{w}} \rangle \simeq \langle \lambda_k | \boldsymbol{w}^\star \rangle$).
- For the less important components $\lambda_k \ll \nu$ we have that $\langle \lambda_k | \tilde{\boldsymbol{w}} \rangle \simeq (\lambda_k/\nu) \langle \lambda_k | \boldsymbol{w}^\star \rangle$ and thus the k-th component of the regularized model gets suppressed.

.

- Observe that:

$$|\boldsymbol{b}\rangle + \boldsymbol{H}\sum_k |\lambda_k\rangle \langle\lambda_k\,|\boldsymbol{w}^\star\rangle = |\boldsymbol{0}\rangle$$
$$\langle\lambda_q\,|\boldsymbol{b}\rangle + \lambda_q \langle\lambda_q\,|\boldsymbol{w}^\star\rangle = 0.$$

- Therefore, for $s_k := \nu/\lambda_k$:

$$E(\boldsymbol{w}^\star) - E_0 = -\frac{1}{2}\sum_k \lambda_k \langle\lambda_k\,|\boldsymbol{w}^\star\rangle^2$$
$$E(\tilde{\boldsymbol{w}}) - E_0 = -\frac{1}{2}\sum_k \lambda_k \frac{1+2s_k}{(1+s_k)^2} \langle\lambda_k\,|\boldsymbol{w}^\star\rangle^2$$
$$\tilde{E}(\boldsymbol{w}^\star) - E_0 = -\frac{1}{2}\sum_k \lambda_k (1-s_k) \langle\lambda_k\,|\boldsymbol{w}^\star\rangle^2$$
$$\tilde{E}(\tilde{\boldsymbol{w}}) - E_0 = -\frac{1}{2}\sum_k \lambda_k \frac{1}{1+s_k} \langle\lambda_k\,|\boldsymbol{w}^\star\rangle^2 ,$$

thus we have that $E(\boldsymbol{w}^\star) < E(\tilde{\boldsymbol{w}})$ but $\tilde{E}(\tilde{\boldsymbol{w}}) < \tilde{E}(\boldsymbol{w}^\star)$.

Regularization: Smoothness Constraint

- Consider the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1}^N$, where $\boldsymbol{x}_n \in \mathscr{X} \subset \mathbb{R}^d$ and $t_n \in \mathbb{R}$. Suppose we have the following loss function:

$$\mathcal{L}[y] = \frac{1}{2}\sum_{n=1}^{N}[y(\boldsymbol{x}_n) - t_n]^2 + \frac{\nu}{2}\int_{\Omega} \mathrm{d}\boldsymbol{x}|\boldsymbol{D}y(\boldsymbol{x})|^2$$

  where $\boldsymbol{D}$ is a suitable differential operator of order p, defined over a region $\Omega \subset \mathbb{R}^d$, and ν is the regularization parameter. Mappings $y : \mathscr{X} \to \mathbb{R}$ with high curvature increase the second term in the error function.

.

- By computing the variation of the error function with respect to $y(\boldsymbol{x})$ we obtain:

$$\begin{aligned}\frac{\delta\mathcal{L}[y]}{\delta y(\boldsymbol{x}_0)} &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{2}\sum_{n=1}^{N}\{[y(\boldsymbol{x}_n)+\lambda\delta(\boldsymbol{x}_n-\boldsymbol{x}_0)]-t_n\}^2+\right.\\ &\left.\left.+\frac{\nu}{2}\int_{\Omega}\mathrm{d}\boldsymbol{x}\left|\boldsymbol{D}\left[y(\boldsymbol{x})+\lambda\delta(\boldsymbol{x}-\boldsymbol{x}_0)\right]\right|^2\right)\right|_{\lambda=0}\\ &=\sum_{n=1}^{N}[y(\boldsymbol{x}_n)-t_n]\,\delta(\boldsymbol{x}_n-\boldsymbol{x}_0)+\nu\hat{\boldsymbol{D}}\boldsymbol{D}y(\boldsymbol{x}_0)=0\end{aligned}$$

where $\hat{\boldsymbol{D}}$ is the operator adjoint to $\boldsymbol{D}$: for all $f, g \in C^{(p)}_{\mathbb{R}^d}$, where $C^{(p)}_{\mathbb{R}^d}$ is the set of functions with at least p derivatives in $\mathbb{R}^d$, we have that:

$$\int_{\mathbb{R}^d}\mathrm{d}\boldsymbol{x}\,f(\boldsymbol{x})\boldsymbol{D}g(\boldsymbol{x})=\int_{\mathbb{R}^d}\mathrm{d}\boldsymbol{x}\,g(\boldsymbol{x})\hat{\boldsymbol{D}}f(\boldsymbol{x}).$$

.

- By considering the Green function $G(\boldsymbol{x}-\boldsymbol{x}')$ of the operator $\hat{\boldsymbol{D}}\boldsymbol{D}$, we can propose the solution as.:

$$\hat{\boldsymbol{D}}\boldsymbol{D}G(\boldsymbol{x}-\boldsymbol{x}') = \delta(\boldsymbol{x}-\boldsymbol{x}')$$

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} w_n G\left(\boldsymbol{x}-\boldsymbol{x}_n\right).$$

Then

$$\sum_{n=1}^{N}\left[y\left(\boldsymbol{x}_n\right)-t_n+\nu w_n\right]\delta(\boldsymbol{x}_n-\boldsymbol{x}_0) = 0$$

$$y\left(\boldsymbol{x}_n\right)-t_n+\nu w_n = 0 \quad \forall 1 \leq n \leq N.$$

.

- If the differential operator is such that:

$$\int_{\Omega} \mathrm{d}x |\boldsymbol{D}y(x)|^2 = \int_a^b \mathrm{d}x \left(|y(x)|^2 + \left| \frac{\mathrm{d}y(x)}{\mathrm{d}x} \right|^2 \right)$$

  and if $y(\partial\Omega) = 0$, variations in the regularization term can be computed as:

$$\begin{aligned}
\delta \int_{\Omega} \mathrm{d}x |\boldsymbol{D}y(x)|^2 &= \int_a^b \mathrm{d}x \left(|y(x) + \delta y(x)|^2 + \left| \frac{\mathrm{d}\left[y(x) + \delta y(x)\right]}{\mathrm{d}x} \right|^2 \right) - \\
&- \int_a^b \mathrm{d}x \left(|y(x)|^2 + \left| \frac{\mathrm{d}y(x)}{\mathrm{d}x} \right|^2 \right) \\
&= 2 \int_a^b \mathrm{d}x \left\{ y(x)\delta y(x) + \frac{\mathrm{d}y(x)}{\mathrm{d}x} \frac{\mathrm{d}\delta y(x)}{\mathrm{d}x} \right\} + O(\delta y^2) \\
&= 2 \ \delta y(x) \frac{\mathrm{d}y(x)}{\mathrm{d}x} \bigg|_a^b + 2 \int_a^b \mathrm{d}x \left(y(x) - \frac{\mathrm{d}^2 y(x)}{\mathrm{d}x^2} \right) \delta y(x).
\end{aligned}$$

- By using that $\frac{\delta y(x)}{\delta y(x_0)} = \delta(x - x_0)$, and that $\delta y(x) \frac{\mathrm{d}y(x)}{\mathrm{d}x}\Big|_a^b = 0$ we have that:

$$\frac{\delta \int_\Omega \mathrm{d}x |\boldsymbol{D}y(x)|^2}{\delta y(x_0)} = 2\hat{\boldsymbol{D}}\boldsymbol{D}y(x_0)$$
$$= \int_a^b \mathrm{d}x \left(1 - \frac{\mathrm{d}^2}{\mathrm{d}x^2}\right) y(x) \frac{\delta y(x)}{\delta y(x_0)},$$

thus

$$\hat{\boldsymbol{D}}\boldsymbol{D}y(x) = \left(1 - \frac{\mathrm{d}^2}{\mathrm{d}x^2}\right) y(x).$$

- Observe the action of this operator on a plane wave:

$$\left(1-\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)\mathrm{e}^{ixs}=\left(1+s^2\right)\mathrm{e}^{ixs}.$$

By considering the Fourier Transform of the Green function we have that

$$\begin{aligned}
G(x-x') &= \int_{-\infty}^{\infty}\frac{\mathrm{d}s}{\sqrt{2\pi}}\tilde{G}(s)\mathrm{e}^{i(x-x')s}\\
\hat{\boldsymbol{D}}\boldsymbol{D}G(x-x') &= \int_{-\infty}^{\infty}\frac{\mathrm{d}s}{\sqrt{2\pi}}\tilde{G}(s)\left(1-\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)\mathrm{e}^{i(x-x')s}\\
&= \int_{-\infty}^{\infty}\frac{\mathrm{d}s}{\sqrt{2\pi}}\tilde{G}(s)\left(1+s^2\right)\mathrm{e}^{i(x-x')s}\\
&= \delta(x-x') = \int_{-\infty}^{\infty}\frac{\mathrm{d}s}{2\pi}\,\mathrm{e}^{i(x-x')s}.
\end{aligned}$$

- Thus

$$\tilde{G}(s) = \frac{1}{\sqrt{2\pi}} \frac{1}{1+s^2}$$

$$G(x-x') = \int_{-\infty}^{\infty} \frac{\mathrm{d}s}{2\pi} \frac{\mathrm{e}^{i(x-x')s}}{1+s^2}$$

$$= \begin{cases} \lim_{R\to\infty} \oint_{C_+(R)} \frac{\mathrm{d}z}{2\pi} \frac{\mathrm{e}^{i(x-x')z}}{1+z^2} & x-x' > 0 \\ \lim_{R\to\infty} \oint_{C_-(R)} \frac{\mathrm{d}z}{2\pi} \frac{\mathrm{e}^{i(x-x')z}}{1+z^2} & x-x' < 0, \end{cases}$$

where $C_\pm(R)$ are the parameterizations presented in the figure.

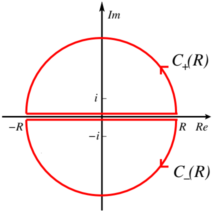


Figure: Contours for the Fourier integral of $\tilde{G}(s)$.

- The integrals are:

$$\begin{aligned}
\lim_{R\to\infty} \oint_{C_\pm(R)} \frac{\mathrm{d}z}{2\pi} \frac{\mathrm{e}^{\pm i|x-x'|z}}{1+z^2} &= \lim_{R\to\infty} \int_{-R}^{R} \frac{\mathrm{d}s}{2\pi} \frac{\mathrm{e}^{\pm i|x-x'|s}}{1+s^2} + \\
&+ \lim_{R\to\infty} \int_0^{\pm\pi} \frac{\mathrm{d}\theta}{2\pi} R \frac{\exp\left(\pm i|x-x'|R\mathrm{e}^{i\theta}\right)}{1+R^2} \\
&= i \sum \mathrm{Res} \left(\frac{\mathrm{e}^{\pm i|x-x'|z}}{1+z^2} \right) .
\end{aligned}$$

- The expansion around the correspondent poles are:

$$\frac{\mathrm{e}^{\pm i|x-x'|z}}{1+z^2} = \frac{\mathrm{e}^{\pm i|x-x'|z}}{(z-i)(z+i)}$$

$$\mathrm{Res}\left(\frac{\mathrm{e}^{\pm i|x-x'|z}}{1+z^2}\right) = \frac{\mathrm{e}^{-|x-x'|}}{2i},$$

therefore

$$\lim_{R\to\infty} \oint_{C_\pm(R)} \frac{\mathrm{d}z}{2\pi} \frac{\mathrm{e}^{\pm i|x-x'|z}}{1+z^2} = \pi \mathrm{e}^{-|x-x'|}$$

$$\left| \int_0^{\pm\pi} \frac{\mathrm{d}\theta}{2\pi} R \frac{\exp\left(\pm i|x-x'|R\mathrm{e}^{i\theta}\right)}{1+R^2} \right| < \int_0^{\pi} \frac{\mathrm{d}\theta}{2\pi} R \frac{\left|\exp\left(\pm i|x-x'|R\mathrm{e}^{i\theta}\right)\right|}{1+R^2}$$
$$< \frac{1}{2}\frac{R}{1+R^2},$$

thus

$$\lim_{R\to\infty} \int_{-R}^{R} \frac{\mathrm{d}s}{2\pi} \frac{\mathrm{e}^{\pm i|x-x'|s}}{1+s^2} = \frac{\mathrm{e}^{-|x-x'|}}{2} = G(x-x').$$

.

- If the differential operator is such that:

$$\int_\Omega \mathrm{d}\boldsymbol{x} |\boldsymbol{D}y(\boldsymbol{x})|^2 = \sum_{\ell=0}^{\infty} \frac{\sigma^{2\ell}}{\ell! 2^\ell} \int_\Omega \mathrm{d}\boldsymbol{x} \left| \mathscr{D}^\ell y(\boldsymbol{x}) \right|^2$$

where: $\mathscr{D}^{2m} = (\nabla^2)^m$ and $\mathscr{D}^{2m+1} = \nabla \mathscr{D}^{2m}$ and if $y(\partial\Omega) = 0$, variations in the regularization term can be computed as:

$$\delta \int_\Omega \mathrm{d}\boldsymbol{x} |\boldsymbol{D}y(\boldsymbol{x})|^2 = \sum_{\ell=0}^{\infty} \frac{\sigma^{2\ell}}{\ell! 2^\ell} \delta \int_\Omega \mathrm{d}\boldsymbol{x} \left| \mathscr{D}^\ell y(\boldsymbol{x}) \right|^2 .$$

- The variation per term are

$$\begin{aligned} \delta \int_\Omega \mathrm{d}\boldsymbol{x} \left| \mathscr{D}^\ell y(\boldsymbol{x}) \right|^2 &= \int_\Omega \mathrm{d}\boldsymbol{x} \left| \mathscr{D}^\ell [y(\boldsymbol{x}) + \delta y(\boldsymbol{x})] \right|^2 - \int_\Omega \mathrm{d}\boldsymbol{x} \left| \mathscr{D}^\ell y(\boldsymbol{x}) \right|^2 \\ &= 2 \int_\Omega \mathrm{d}\boldsymbol{x} \mathscr{D}^\ell y(\boldsymbol{x}) \mathscr{D}^\ell \delta y(\boldsymbol{x}) + O(\delta y^2) \end{aligned}$$

- If $\ell = 2m$ we have, disregarding terms of $O(\delta y^2)$:

$$\begin{aligned}
\delta \int_\Omega \mathrm{d}\boldsymbol{x} \left|\mathscr{D}^{2m} y(\boldsymbol{x})\right|^2 &= 2 \int_\Omega \mathrm{d}\boldsymbol{x} \left(\mathscr{D}^{2m} y(\boldsymbol{x})\right) \; \nabla^2 (\nabla^2)^{m-1} \delta y(\boldsymbol{x}) \\
&= 2 \int_\Omega \mathrm{d}\boldsymbol{x} \left(\mathscr{D}^{2m} y(\boldsymbol{x})\right) \; \left\langle \nabla \,\middle|\, \nabla (\nabla^2)^{m-1} \delta y(\boldsymbol{x}) \right\rangle \\
&= 2 \int_\Omega \mathrm{d}\boldsymbol{x} \left(\mathscr{D}^{2m} y(\boldsymbol{x})\right) \; \left\langle \nabla \,\middle|\, \mathscr{D}^{2m-1} \delta y(\boldsymbol{x}) \right\rangle \\
&= 2 \, \mathscr{D}^{2m} y(\boldsymbol{x}) \left\langle \mathscr{D}^{2m-1} \delta y(\boldsymbol{x}) \,\middle|\, \boldsymbol{n}_{\partial\Omega} \right\rangle \Big|_{\partial\Omega} - \\
&- 2 \int_\Omega \mathrm{d}\boldsymbol{x} \left\langle \nabla \mathscr{D}^{2m} y(\boldsymbol{x}) \,\middle|\, \mathscr{D}^{2m-1} \delta y(\boldsymbol{x}) \right\rangle ,
\end{aligned}$$

where $\boldsymbol{n}_{\partial\Omega}$ is the unit vector perpendicular to the closure $\partial\Omega$ of the volume considered Ω.

- If $y(\boldsymbol{x})$ and its derivatives evaluate to zero in this boundary, the first term is zero, then

$$\begin{aligned}
\delta \int_\Omega \mathrm{d}\boldsymbol{x} \left|\mathscr{D}^{2m} y(\boldsymbol{x})\right|^2 &= -2 \int_\Omega \mathrm{d}\boldsymbol{x} \left\langle \mathscr{D}^{2m+1} y(\boldsymbol{x}) \right| \mathscr{D}^{2m-1} \delta y(\boldsymbol{x}) \big\rangle \\
&= -2 \int_\Omega \mathrm{d}\boldsymbol{x} \left\langle \mathscr{D}^{2m+1} y(\boldsymbol{x}) \right| \nabla \mathscr{D}^{2m-2} \delta y(\boldsymbol{x}) \big\rangle \\
&= -2 \left\langle \mathscr{D}^{2m+1} y(\boldsymbol{x}) \right| \boldsymbol{n}_{\partial\Omega} \big\rangle \, \mathscr{D}^{2m-2} \delta y(\boldsymbol{x}) \big|_{\partial\Omega} + \\
&+ 2 \int_\Omega \mathrm{d}\boldsymbol{x} \left(\mathscr{D}^{2m+2} y(\boldsymbol{x}) \right) \; \mathscr{D}^{2m-2} \delta y(\boldsymbol{x}),
\end{aligned}$$

and repeating for $2m$ times we obtain:

$$\delta \int_\Omega \mathrm{d}\boldsymbol{x} \left|\mathscr{D}^{2m} y(\boldsymbol{x})\right|^2 = 2 \int_\Omega \mathrm{d}\boldsymbol{x} \mathscr{D}^{4m} y(\boldsymbol{x}) \, \delta y(\boldsymbol{x}).$$

- If $\ell = 2m+1$ we have that:

$$\begin{aligned}\delta \int_\Omega \mathrm{d}\boldsymbol{x} \left|\mathscr{D}^{2m+1} y(\boldsymbol{x})\right|^2 &= 2\int_\Omega \mathrm{d}\boldsymbol{x} \left\langle \mathscr{D}^{2m+1} y(\boldsymbol{x}) \middle| \mathscr{D}^{2m+1} \delta y(\boldsymbol{x})\right\rangle \\ &= 2 \left\langle \mathscr{D}^{2m+1} y(\boldsymbol{x}) \middle| \boldsymbol{n}_{\partial\Omega}\right\rangle \mathscr{D}^{2m} \delta y(\boldsymbol{x})\Big|_{\partial\Omega} - \\ &- 2\int_\Omega \mathrm{d}\boldsymbol{x} \left(\mathscr{D}^{2m+2} y(\boldsymbol{x})\right) \ \mathscr{D}^{2m} \delta y(\boldsymbol{x}),\end{aligned}$$

and the rest of the argument follows the same path until:

$$\delta \int_\Omega \mathrm{d}\boldsymbol{x} \left|\mathscr{D}^{2m+1} y(\boldsymbol{x})\right|^2 = -2 \int_\Omega \mathrm{d}\boldsymbol{x} \mathscr{D}^{4m+2} y(\boldsymbol{x}) \delta y(\boldsymbol{x}).$$

- Thus

$$\delta \int_\Omega \mathrm{d}\boldsymbol{x} \left|\mathscr{D}^{\ell} y(\boldsymbol{x})\right|^2 = 2(-1)^\ell \int_\Omega \mathrm{d}\boldsymbol{x} \mathscr{D}^{2\ell} y(\boldsymbol{x})\, \delta y(\boldsymbol{x}).$$

- By using that $\frac{\delta y(\boldsymbol{x})}{\delta y(\boldsymbol{x}_0)} = \delta(\boldsymbol{x} - \boldsymbol{x}_0)$,we have that

$$\begin{aligned}
\frac{\delta}{\delta y(\boldsymbol{x}_0)} \int_\Omega \mathrm{d}\boldsymbol{x} |\boldsymbol{D} y(x)|^2 &= 2\hat{\boldsymbol{D}}\boldsymbol{D} y(\boldsymbol{x}_0) \\
&= \sum_{\ell=0}^{\infty} \frac{\sigma^{2\ell}}{\ell! 2^\ell} \frac{\delta}{\delta y(\boldsymbol{x}_0)} \int_\Omega \mathrm{d}\boldsymbol{x} \left| \mathscr{D}^\ell y(\boldsymbol{x}) \right|^2 \\
&= 2 \sum_{\ell=0}^{\infty} \frac{\sigma^{2\ell}}{\ell! 2^\ell} (-1)^\ell \int_\Omega \mathrm{d}\boldsymbol{x} \mathscr{D}^{2\ell} y(\boldsymbol{x}) \frac{\delta y(\boldsymbol{x})}{\delta y(\boldsymbol{x}_0)} \\
&= 2 \sum_{\ell=0}^{\infty} \frac{\sigma^{2\ell}}{\ell! 2^\ell} (-1)^\ell \mathscr{D}^{2\ell} y(\boldsymbol{x}_0)
\end{aligned}$$

Thus

$$\hat{\boldsymbol{D}}\boldsymbol{D} y(\boldsymbol{x}) = \sum_{\ell=0}^{\infty} \frac{\sigma^{2\ell}}{\ell! 2^\ell} (-1)^\ell \mathscr{D}^{2\ell} y(\boldsymbol{x}).$$

.

Observe that the Laplacian of a plane wave is:

$$\nabla^2 \mathrm{e}^{i\langle \boldsymbol{x}|\boldsymbol{s}\rangle} = -\left\langle \boldsymbol{s}\middle|\, \boldsymbol{s}\right\rangle \mathrm{e}^{i\langle \boldsymbol{x}|\boldsymbol{s}\rangle}.$$

By considering the Fourier Transform of the Green function we have that

$$\begin{aligned}
G(\boldsymbol{x}-\boldsymbol{x}') &= \int \frac{\mathrm{d}\boldsymbol{s}}{(2\pi)^{d/2}}\,\tilde{G}(\boldsymbol{s})\mathrm{e}^{i\langle \boldsymbol{x}-\boldsymbol{x}'|\boldsymbol{s}\rangle} \\
\hat{\boldsymbol{D}}\boldsymbol{D}G(\boldsymbol{x}-\boldsymbol{x}') &= \int \frac{\mathrm{d}\boldsymbol{s}}{(2\pi)^{d/2}}\,\tilde{G}(\boldsymbol{s})\sum_{\ell=0}^{\infty}\frac{\sigma^{2\ell}}{\ell!2^{\ell}}(-1)^{\ell}\mathscr{D}^{2\ell}\mathrm{e}^{i\langle \boldsymbol{x}-\boldsymbol{x}'|\boldsymbol{s}\rangle} \\
&= \int \frac{\mathrm{d}\boldsymbol{s}}{(2\pi)^{d/2}}\,\tilde{G}(\boldsymbol{s})\sum_{\ell=0}^{\infty}\frac{\sigma^{2\ell}}{\ell!2^{\ell}}\left\langle \boldsymbol{s}\middle|\, \boldsymbol{s}\right\rangle^{\ell}\mathrm{e}^{i\langle \boldsymbol{x}-\boldsymbol{x}'|\boldsymbol{s}\rangle} \\
&= \int \frac{\mathrm{d}\boldsymbol{s}}{(2\pi)^{d/2}}\,\tilde{G}(\boldsymbol{s})\exp\left(\frac{1}{2}\left\langle \boldsymbol{s}\,\middle|\sigma^2\boldsymbol{I}\,\middle|\,\boldsymbol{s}\right\rangle\right)\mathrm{e}^{i\langle \boldsymbol{x}-\boldsymbol{x}'|\boldsymbol{s}\rangle} \\
&= \delta(\boldsymbol{x}-\boldsymbol{x}') = \int \frac{\mathrm{d}\boldsymbol{s}}{(2\pi)^{d}}\,\mathrm{e}^{i\langle \boldsymbol{x}-\boldsymbol{x}'|\boldsymbol{s}\rangle}.
\end{aligned}$$

Radial Basis Functions

Thus

$$\tilde{G}(\boldsymbol{s}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\left\langle \boldsymbol{s} \,\middle|\sigma^2 \boldsymbol{I}\middle|\, \boldsymbol{s}\right\rangle\right)$$
$$G(\boldsymbol{x} - \boldsymbol{x}') = \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}', \sigma^2 \boldsymbol{I}).$$

- Then $y(\boldsymbol{x}) = \sum_{n=1}^{N} w_n \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$.

By defining the matrix $\boldsymbol{G}$ with entries $[\boldsymbol{G}]_{nm} = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{x}_m, \sigma^2 \boldsymbol{I})$, and the vectors $[\boldsymbol{w}]_n = w_n$ and $[\boldsymbol{t}]_n = t_n$ we have that the network weights are the solutions to the linear equation

$$(\boldsymbol{G} + \nu \boldsymbol{I})\boldsymbol{w} = \boldsymbol{t}.$$

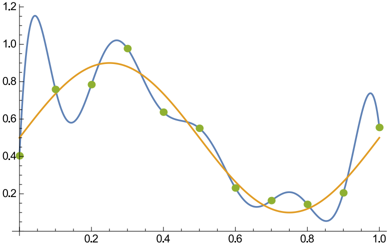
.

Application: Interpolation

- Each input $\boldsymbol{x}_n$ must be mapped onto an output t_n so that $y(\boldsymbol{x}_n) = t_n$.
- $y(\boldsymbol{x}_n)$ is obtained by linear superposition of the same N basis functions $y(\boldsymbol{x}_n) = \sum_{m=1}^{N} w_m \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_m, \sigma^2 \boldsymbol{I})$.

The parameters w_n are obtained by $w_n = \sum_m [\boldsymbol{G}^{-1}]_{nm} t_m$ where $\boldsymbol{G}^{-1}$ is the inverse matrix with entries $[\boldsymbol{G}]_{n,m} = \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_m, \sigma^2 \boldsymbol{I})$.

- Where $N = 11$ and $t_n = 0.5 + 0.4\sin(2\pi x_n) + \epsilon_n$ with $\epsilon_n \sim \mathcal{N}(\epsilon|0, 0.1)$.
- The basis functions are Gaussians with variance $\sigma^2 = 0.065$.
- In the graph, the green dots are the data points, the orange curve represents the generator function (deterministic part), and in blue is the interpolation curve.

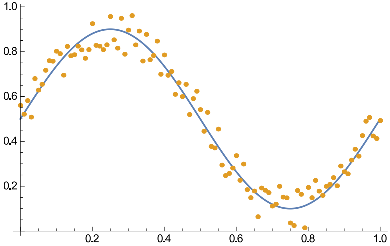


RBN: Modified interpolation

1. The number K of basis functions need not equal the number N of data points, and is typically much less than N.
2. The centres of the basis functions are no longer constrained to be given by input data vectors. Instead, the determination of suitable centres becomes part of the training process.

The update algorithm for the weights can be constructed based on a sum of squares error function.

.

- The prototype and other parameters for the kernels ϕ are computed from the independent part of the data points (unsupervised learning).
- The weights $\boldsymbol{w}$ are obtained afterwards by gradient descent.
- Example: Let us use the data generator presented above to construct a data set with 100 points.
- 

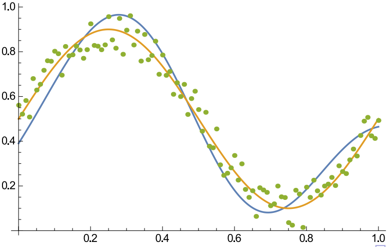
- The centers $\boldsymbol{\mu}_k$ are obtained by:

$$\boldsymbol{\mu}_k = \frac{K}{N} \sum_{\ell=1}^{K/N} \mathbf{x}_{\ell+(k-1)K}.$$

- The kernels $\phi_k(\mathbf{x})$ are Gaussians centered at $\boldsymbol{\mu}_k$ with variance σ^2 (same for all).
- The algorithm for the weights is:

$$\mathbf{w}_{m+1} = \mathbf{w}_m - \eta[y(\mathbf{x}_m) - t_m]\phi(\mathbf{x}_m)$$

where η is the learning rate and $[\phi]_k = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \sigma^2 \mathbf{I})$.

- 
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0