Noise Injection: theoretical prospects

Yves Grandvalet[1], Stéphane Canu[1] and Stéphane Boucheron[2]
(1) Heudiasyc, U.R.A. C.N.R.S. 817
Université de Technologie de Compiègne
Centre de Recherches de Royallieu
B.P. 529, 60205 Compiègne Cedex, France

grandval@hds.univ-compiegne.fr
scanu@hds.univ-compiegne.fr

(2) LRI-CNRS
Université Paris-Sud, Bât. 490
91405 Orsay Cedex, France

bouchero@lri.fr

September 17, 1996

Abstract

Noise Injection consists in adding noise to the inputs during neural network training. Experimental results suggest that it might improve the generalization ability of the resulting neural network. A justification of this improvement remains elusive: first, describing analytically the average perturbed cost function is difficult, second, controlling the fluctuations of the random perturbed cost function is hard. Hence recent papers suggest to replace the random perturbed cost by a (deterministic) Taylor approximation of the average perturbed cost function. This paper takes a different stance: when the injected noise is Gaussian, Noise Injection is naturally connected to the action of the Heat Kernel. This provides indications on the relevance domain of traditional Taylor expansions, and shows the dependence of the quality of Taylor approximations on global smoothness properties of neural networks under consideration. The connection between noise injection and Heat kernel also enables to control the fluctuations of the random perturbed cost function. Under the previously mentioned global smoothness assumption, tools from Gaussian analysis provide bounds on the tail behavior of the perturbed cost. This finally suggests that mixing input perturbation with smoothness based penalization might be profitable.

1 Introduction

1.1 Noise Injection

Neural network training consists in minimizing a cost functional $C(.)$ on the set of functions $\mathcal{F}$ realizable by multi-layer perceptrons (MLP) with fixed architecture. The cost C is usually the averaged squared error:

$$C(f) = \mathbb{E}_Z \left(f(X) - Y \right)^2 \tag{1}$$

where the random variable $Z = (X, Y)$ describing the data is sampled according to a fixed but unknown law. In applications, as C is not computable, an empirically computable cost is then minimized using a sample $\mathbf{z}_\ell = \{\mathbf{z}^i\}_{i=1}^\ell$, with $\mathbf{z}^i = (\mathbf{x}^i, y^i) \in \mathbb{R}^d \times \mathbb{R}$ gathered by drawing independent identically-distributed data according to the law of Z. An estimate $\hat{f}_{emp}$ of the regression function $f^*(\mathbf{x}) = \arg\min_{f \in L^2} C(f)$, is given by the minimization of the empirical cost C_{emp}:

$$C_{emp}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mathbf{x}^i) - y^i \right)^2 \tag{2}$$

The cost $C_{emp}(.)$ is a random functional with expectation $C(.)$. In order to justify the minimization of C_{emp}, the convergence of the empirical cost towards its expectation should be uniform with respect to $f \in \mathcal{F}$ [Vapnik, 1982, Haussler, 1992]. When $\mathcal{F}$ is too large, this may not hold against some sampling laws. Hence practitioners have to trade the expressive power of $\mathcal{F}$ with the ability to control the fluctuations of $C_{emp}(.)$.

This suggests analyzing modified estimators that possibly restrict the effective search space. One of these modified training methods consists in applying perturbations to the inputs during training. Experimental results in [Sietsma and Dow, 1991] show that Noise Injection (NI) can dramatically improve the generalization ability of MLP. This is especially attractive because the modified minimization problem can be solved thanks to the initial training algorithm. During NI, the original training sample $\mathbf{z}_\ell$ is distorted by adding some noise $\boldsymbol{\eta}$ to the inputs $\mathbf{x}^i$ while leaving the target value y^i unchanged. During the k^{th} epoch of the back-propagation algorithm, a new distortion $\boldsymbol{\eta}^k$ is applied to $\mathbf{z}_\ell$. The distorted sample is then used to compute the error and to derive the weights updates. A stochastic algorithm is thus obtained that eventually minimizes $C_{NI^m_{emp}}$, defined as:

$$C_{NI^m_{emp}}(f) = \frac{1}{m} \sum_{k=1}^{m} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mathbf{x}^i + \boldsymbol{\eta}^{k,i}) - y^i \right)^2 \tag{3}$$

where the number of replications m is set under user control but finite. The average value of the perturbed cost is:

$$C_{NI}(f) = \mathbb{E}_{\boldsymbol{\eta}} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mathbf{x}^i + \boldsymbol{\eta}) - y^i \right)^2 \right] \tag{4}$$

In this paper, the noise $\boldsymbol{\eta}$ is assumed to be a centered Gaussian vector with independent coordinates: $\mathbb{E}[\boldsymbol{\eta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\eta}^T \boldsymbol{\eta}] = \sigma^2 \mathbf{I}$.

The success of NI is intuitively explained by asserting that minimizing C_{NI} (4) ensures that similar inputs lead to similar outputs. It raises two questions: when should we prefer to minimize C_{NI} rather than C_{emp}? How does $C_{NI^m_{emp}}$ converge towards C_{NI}?

Recently, several papers [Webb, 1994, Bishop, 1995, Leen, 1995, Reed et al., 1995, An, 1995] have resorted to Taylor expansions to describe the impact of NI, and to motivate the minimization of C_{NI} rather than C_{emp}. Those papers not only try to provide a formal description of NI but also aim at finding a deterministic alternative to the minimization of $C_{NI^m_{emp}}$.

1.2 Organization of the paper

This paper takes a different approach: when the injected noise is Gaussian, the Taylor expansion approach is connected to the action of the Heat Kernel, and the dependance of C_{NI} (4) on the noise variance is shown to obey the Heat equation (Section 2.1). This clear connection between partial differential equations and NI provides some indications on the relevance domain of traditional Taylor expansions (Section 2.2). Finally, we analyze the simplified expressions that are assumed to be valid locally around optimal solutions (Section 2.3).

The connection between NI and the action of the Heat kernel also enables to control the fluctuations of the random perturbed cost function. Under some natural global smoothness property of the class of multi-layer perceptrons under consideration, tools from Gaussian analysis provide exponential bounds on the probability of deviation of the perturbed cost (Section 3.3). This finally suggests that mixing NI with smoothness based penalization might be profitable (Section 3.4).

2 Taylor expansions

2.1 Gaussian perturbation and Heat equation

To exhibit the connection between Gaussian Noise Injection and the Heat equation, let us define u as a function from $\mathbb{R}^+ \times \mathbb{R}^{\ell \times d}$ by:

$$u(0, \mathbf{x}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \Big(f(\mathbf{x}^i) - y^i \Big)^2 \tag{5}$$

$$u(t, \mathbf{x}) = \mathbb{E}_{\boldsymbol{\eta}} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} \Big(f(\mathbf{x}^i + \boldsymbol{\eta}^i) - y^i \Big)^2 \right] \tag{6}$$

Obviously, we have $u(0, \mathbf{x}) = C_{emp}(f)$ and $u(t, \mathbf{x}) = C_{NI}(f)$, when t is the noise variance σ^2. Each value of the noise variance defines a linear operator T_t that maps $u(0, .)$ onto $u(t, .)$. Moreover since the sum of two independent Gaussian with variances s and t is Gaussian with variance $s + t$, the family $\big(T_t\big)_{t \geq 0}$ defines a semigroup, the *Heat semigroup* (cf. [Ethier and Kutrz, 1986] for an introduction to semigroup operators). The function u obeys the Heat equation (cf. [Karatzas and Shreve, 1988, Chapter 4, Sections 3 and 4]):

$$\frac{\partial u}{\partial t} = \frac{1}{2} \boldsymbol{\Delta}_{\mathbf{xx}} u \tag{7}$$

where $\boldsymbol{\Delta}_{\mathbf{xx}}$ is the Laplacian w.r.t. $\mathbf{x}$, and where the initial conditions are defined by (5). For the sake of self-containment, a derivation of equation (7) is given in the appendix when initial conditions are square-integrable. Let us denote by $C_{NI}(f, t)$ the perturbed cost when the noise is Gaussian of variance is t, eq. (7) yields:

$$C_{NI}(f, t) = C_{emp}(f) + \frac{1}{2} \int_0^t \boldsymbol{\Delta}_{\mathbf{xx}} C_{NI}(f, s) \; ds \;\; . \tag{8}$$

Therefore, C_{NI} can be investigated in the purely analytical framework of partial differential equations (under the Gaussian assumption). The possibility to forget about the original probabilistic setting when dealing with neural networks follows from the Tikhonov uniqueness theorem [Karatzas and Shreve, 1988, Chapter 4, Section 4.3].

Observation 2.1 *If $\mathcal{F}$ is a class of functions definable by some feedforward architecture using sigmoidal, radial basis functions, or piecewise polynomials as activation functions, and if injected noise follows a Gaussian law, then the perturbed cost C_{NI} is the unique function of the variance t that obeys the Heat equation (7) with initial conditions defined in (5).*

Any deterministic faithful simulation of NI should use some numerical analysis software to integrate the Heat equation and then run some back-propagation

software on the result. We do not recommend such a methodology for efficiency reason and insist on the fact that stochastic representations of solutions of PDEs have proved useful in analysis [Karatzas and Shreve, 1988].

One should notice that methods reported in the literature (finite differences, finite element, ... cf. [Press et al., 1992]) assume that the function defining the initial conditions has bounded support. This assumption that makes sense in Physics is not valid in the neural network setting. *Hence Monte-Carlo methods appear to be ideal technique to solve the PDE problem raised by NI.*

2.2 Taylor expansion validity domain

Let $C_{Taylor}(f)$ be the first order Taylor expansion of $C_{NI}(f)$ as a function of $t = \sigma^2$. For various kinds of noise and function classes $\mathcal{F}$, it has been shown in [Matsuoka, 1992, Webb, 1994, Grandvalet and Canu, 1995, Bishop, 1995, Reed et al., 1995] that:

$$C_{Taylor}(f) \quad = \quad C_{emp}(f) + \frac{\sigma^2}{2} \, \mathbf{\Delta_{xx}} C_{emp}(f) \tag{9}$$

In the context of Gaussian Noise Injection, this just means that the Laplacian is the *infinitesimal generator* of the Heat semigroup. To emphasize the distinction between (9) and (7), one should stress the fact that *the Heat equation is not only a correct description of the impact of Gaussian Noise Injection in the small variance limit but also for any value of the variance.*

The Taylor approximation validity domain is restricted to those functions f such that

$$\lim_{\sigma^2 \to 0} \frac{C_{NI}(f) - C_{Taylor}(f)}{\sigma^2} = 0 \ . \tag{10}$$

Observation 2.2 *A sufficient condition for the Taylor approximation to be valid is that C_{emp} belongs to the domain of the generator of the Heat semigroup.*

The empirical cost C_{emp} has to be a licit initial condition for the Heat equation, which is always true in the neural network context (cf. conditions in [Karatzas and Shreve, 1988, Theorems 3.3 and 4.2, Chapter 4]).

The preceding statement is purely analytical and does not say much about the relevance of minimizing C_{Taylor} while training neural networks. This issue may be analyzed according to several directions: is the minimization of C_{Taylor} equivalent to the minimization of C_{NI} ? Is the minimization of C_{Taylor} interesting in its own right ?

The second issue and related developments are addressed in [Bishop, 1995, Leen, 1995]. The first issue can not be settled in a positive way for arbitrary variances in general, but the principle of minimizing C_{NI} (and C_{Taylor}) should be definitively ruled out if if the minima of C_{NI} (and C_{Taylor}) did not converge towards the minima of C_{emp} when $t = \sigma^2 \to 0$. This cannot be deduced directly from (10) since (10) only describes simple convergence.

This will take place when the convergence in equation (10) is uniform over $\mathcal{F}$. As we will vary t, let us denote by $C_{NI}(f,t)$ (resp. $C_{Taylor}(f,t)$) the perturbed cost (resp. its Taylor approximation) of f when the noise variance is t, we get:

$$C_{NI}(f,t) - C_{Taylor}(f,t) = \frac{1}{2}\int_0^t \left[\mathbf{\Delta_{xx}} C_{NI}(f,s) - \mathbf{\Delta_{xx}} C_{emp}(f)\right] ds \tag{11}$$

If some uniform (over $\mathcal{F}$ and $s \leq t_0 < 0$) bound on $|\mathbf{\Delta_{xx}} C_{NI}(f,s) - \mathbf{\Delta_{xx}} C_{emp}(f)|$ is available, then $\lim_{t\to 0} \max_{f\in\mathcal{F}} C_{NI}(f,t) - C_{Taylor}(f,t) = 0$ and little manipulations using the triangular inequality reveal the convergence of minima. The same argument shows the convergence of the minima of C_{NI} towards minima of C_{emp}. If some upper bounds is imposed on the weights of a sigmoidal neural network, those global bounds are automatically enforced.

Imposing bounds on weights is an important requirement to insure the validity of the Taylor approximation. We intuitively expect the truncation of the Taylor series to be valid in the small variance limit. But if $f(x) = g(wx)$, where g is a parameterized function and w is a free parameter, then $f(x+\eta) = g(wx+w\eta)$. The noise $\boldsymbol{\eta}$ injected in f appears to g as a noise $w\boldsymbol{\eta}$, that is, as a noise of variance $w^2\sigma^2$. Therefore, if g is non-linear (i.e. f non-linear), the Taylor expansion will fail for large w^2.

A simple illustration is given on fig. 1. The sample contains 10 (x,y) pairs, where the x^i are regularly placed on $[-0.5, 0.5]$, and $y = 1_{x<0}$: $\mathbf{z}_\ell = \{(-0.5,1),(-0.4,1),\ldots,(-0.1,1),(0.1,0),\ldots,(0.5,0)\}$. The class $\mathcal{F}$ is defined by $f(x) = \exp(ax+b)/(1+\exp(ax+b))$ $a,b \in \mathbb{R}$, and the quadratic loss is used. The three error surfaces for C_{emp}, C_{Taylor}, and C_{NI} are represented for $\sigma^2 = 2.5\ 10^{-3}$. Although the noise variance is quite small, there are some noticeable differences: the cost C_{NI} is smoother than C_{emp} (effect of convolution), which is in turn smoother than C_{Taylor}. The dissimilarities of these surfaces can be shown for any non-zero value of σ^2, by a proper choice of the scale. The bottom right drawings show the error as a function of b for two values of the scale parameter a. It is shown how the Taylor expansion becomes more and more innacurate, as a^2 increases. Note that the number of oscillations on C_{Taylor} is equal to the number of the! points in the sample. Increasi

Remark. Although (9) has been established for various kind of noise (strong integrability properties are sufficient), C_{NI} is the solution of the Heat equation only for Gaussian noise. In non Gaussian contexts NI usually does not define a semigroup of operators indexed by variance.

2.3 Local analysis and simplified expressions

Requiring uniform bounds on $|\mathbf{\Delta_{xx}} C_{NI}(f,s) - \mathbf{\Delta_{xx}} C_{emp}(f)|$ may be too demanding. Fortunately, it is possible to adapt the *local approach* to Noise Injection suggested in [Leen, 1995, Bishop, 1995] to provide another derivation of the behavior of the minima of C_{NI} and C_{Taylor}. The local approach is concerned with the behavior of C_{NI}, C_{Taylor} and possibly other cost functions such as the

derivative-based regularizer [Leen, 1995, Bishop, 1995] around minima (global or local) of C_{emp}. Though Bishop and Leen focused their attention on the case where $\mathbb{E}(Y|X=\mathbf{x})$ is realizable by the neural architecture, and where the empirical measure closely mimics the sampling law (cf. for example [Leen, 1995, Section 3.1.1.]), we believe that their local approach can be extended and applied to the comparison of the *critical points* of $C_{emp}, C_{NI},$ and C_{Taylor}.

A critical point of C_{emp} (and similarly for other costs) is a weight assignment where the gradient of C_{emp} vanishes. A critical point is *non-degenerate* iff the Hessian matrix has full rank. In the sequel, we will assume that C_{emp} has only non degenerate critical points. This is not a major restriction since the set of target values y^i for which C_{emp} does have degenerate critical points has measure 0 for the kind of neural architectures mentioned here. Moreover the number of non degenerate critical points is finite [Sontag, 1995].

Observation 2.3 *If C_{emp} has no degenerate critical points, then there exists some $t_0 > 0$ such that for any $t, 0 \leq t \leq t_0$, to any critical point of C_{emp}, there correspond a critical point of C_{Taylor} and a critical point of C_{NI}, moreover those critical points are within distance Kt of each other for some constant K that depends on the sample under consideration.*

Proof. The argument extends Leen's suggestion ([Leen, 1995]). In the course of the argument we will assume that C_{NI} and C_{Taylor} have partial derivatives with respect to t at $t = 0$, this can be enforced by taking a linear continuation for $t < 0$.

Let us assume that $\mathcal{F}$ is parameterized by W weights. Let $\nabla_{\mathbf{w}} C_{NI}(f,t)$ and $\nabla_{\mathbf{w}} C_{Taylor}(f,t)$ denote the gradient of C_{NI} and C_{Taylor} with respect to the weight assignment for some value of f and $\sigma^2 = t$ (note that at $t = 0$ the two values coincide with $\nabla_{\mathbf{w}} C_{emp}(f)$). If $f^\bullet$ is some non degenerate critical point of C_{emp} then the matrix of partial derivatives of $\nabla_{\mathbf{w}} C_{NI}(f,t)$ and $\nabla_{\mathbf{w}} C_{Taylor}(f,t)$ with respect to weights and time has full rank, thus by the *implicit function theorem in its surjective form* [Hirsch, 1976, page 214], there exists a neighborhood of $(f^\bullet, 0)$, and diffeomorphisms ϕ and ψ defined in a neighborhood of $(\mathbf{0}, 0) \in \mathbb{R}^{W+1}$, such that $\phi(\mathbf{0}, 0) = (f^\bullet, 0)$ (resp. $\psi(\mathbf{0}, 0) = (f^\bullet, 0)$) and $\nabla_{\mathbf{w}} C_{NI}(\phi(\mathbf{u}, v)) = \mathbf{u}$ (resp. $\nabla_{\mathbf{w}} C_{Taylor}(\psi(\mathbf{u}, v)) = \mathbf{u}$). The gradients of ϕ and ψ at $(\mathbf{0}, 0)$ are of L^2 norm less or equal than the norm of the matrix of partial derivatives of $\nabla_{\mathbf{w}} C_{emp}(f^\bullet, 0)$. For sufficiently small values of t, $\phi(\mathbf{0}, t)$ and $\psi(\mathbf{0}, t)$ define continuous curves of critical points of $C_{NI}(., t)$ and $C_{Taylor}(., t)$. As the number of critical points of C_{emp} is finite, we may assume that those curves do not intersect, and that the norm of the gradients of $\phi(\mathbf{0}, t)$ and $\psi(\mathbf{0}, t)$ with respect to t are upper-bounded. The observation follows □

Remark 1. The observation asserts that for sufficiently small t, the local minima of C_{emp} can be injected in the set of local minima of C_{Taylor} and C_{NI}. For C_{Taylor} the reverse is true. The definition of C_{Taylor} obeys the same constraints as the definition of C_{emp}: it is defined using solely $+, \times$, constants and exponentiation, hence, by Sontag bound, for any t, except on a set of measure 0 of

target values y, the number of critical points of C_{Taylor} is finite, and the argument that was used in the proof works in the other direction. For sufficiently small t, C_{Taylor} does not introduce new local minima. This argument cannot be adapted to C_{NI} which definition also requires an integration. Nevertheless, experimental results suggest that NI tends to suppress spurious local minima [Grandvalet, 1995].

Remark 2. The validity of observation relies on the choice of activation functions. If $\mathcal{F}$ were constituted by the class of functions parameterized by $\alpha \geq 0$ mapping $x \mapsto \sin(\alpha x)$. One could manufacture a sample such that C_{emp} and C_{Taylor} both have countably many non degenerate minima, which are in one-to-one correspondence and such that the convergence of the minima of C_{NI} towards the minima of C_{emp} as t tends towards 0 is not uniform.

3 Noise injection and generalization

The alleged improvement in generalization provided by NI remains to be analytically confirmed and explained. Most attempts to provide with an explanation resort to the concepts of *penalization* and *regularization*. Usually penalization consists in adding a positive functional $\Omega(.)$ to the empirical risk. It is called a regularization if the sets $\left\{ f \ : \ f \in \mathcal{F}, \quad \Omega(f) \leq \alpha \right\}$ are compact for the topology on $\mathcal{F}$. Penalization and regularization are standard ways of improving regression and estimation techniques (cf. for example [Grenander, 1981, Vapnik, 1982, Barron et al., 1995]).

When $\mathcal{F}$ is the set of linear functions, Noise Injection has been recognized has a *regularizer in the Tikhonov sense* [Tikhonov and Arsenin, 1977]. This is at present time the only reported case. It is enough to consider $\mathcal{F}$ as constituted by univariate degree 2 polynomial to realize that NI can not generally be regarded as a *penalization* procedure [Barron et al., 1995]: $C_{NI} - C_{emp}$ is not always positive. Thus, the improvement of generalization ability attributed to NI still requires some explanations.

3.1 Noise injection and kernel density estimation

An appealing interpretation connects Noise Injection with kernel estimation techniques [Comon, 1992, Holmström and Koistinen, 1992, Webb, 1994].

Minimization of the empirical risk might be a poor or inefficient heuristic because the minima (if there are any) of C_{emp} could be far away from those of C. Recall that when trying to perform regression with sigmoidal neural networks, we have no guarantees that C_{NI} has a single global minima, or even that the infimum of C_{NI} is realized [Auer et al., 1996]. Hence the safest (and up to the authors knowledge, only) way to warrant the consistency of the NI technique is to get a global control on the fluctuations of $C_{NI}(.)$ with respect to $C(.)$ i.e. on $\sup_{\mathcal{F}} |C_{NI}(f) - C(f)|$. The bad performance of minimization of empirical risk could be due to the slow convergence of the empirical measure

$\hat{p}_Z \stackrel{\Delta}{=} \sum_i \delta_{\mathbf{x}^i, y^i}$ towards the sampling probability in the pseudo-metric induced by $\mathcal{F}$ ($d_{\mathcal{F}}(\hat{p} - \hat{p}') \stackrel{\Delta}{=} \sup_{f \in \mathcal{F}} |\mathbb{E}_{\hat{p}}(f(X) - Y)^2 - \mathbb{E}_{\hat{p}'}(f(X) - Y)^2|$).

But minimizing C_{NI} is equivalent (up to an irrelevant constant factor) to minimizing

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \mathbb{E}_{\boldsymbol{\eta}, \boldsymbol{\eta}'} \left(f(\mathbf{x}^i + \boldsymbol{\eta}^i) - (y^i + \boldsymbol{\eta}') \right)^2, \tag{12}$$

where $\boldsymbol{\eta}'$ is a scalar Gaussian independent of $\boldsymbol{\eta}$. Minimizing C_{NI} consists in minimizing the empirical risk against a smoothed version of the empirical measure: the Parzen-Rosenblatt estimator of the density (cf. for details on the latter, [Devroye, 1987]). The connection with Gaussian kernel density estimation and regularization can thus be established in a perspective described in [Grenander, 1981]. As the Gaussian kernel is the fundamental solution of the Heat equation, the smoothed density that defines (12) is obtained by running the Heat equation using the empirical measure as an initial condition (this is called the Weierstrass transform in [Grenander, 1981]). It is then tempting to explain the improvement in generalization provided by NI using upper bounds on the rate of convergence of the Parzen-Rosenblatt estimator. Though conceptually appealing, this approach might be disappointing: it actually suggests to solve a density estimation problem as a subproblem of a regression problem. The former is much harder than the latter [Vapnik, 1982, Devroye, 1987].

3.2 Consistency of NI

To assess the consistency of minimizing C_{NI}, we would like to control $C_{NI}(.) - C(.)$. A reasonable way of analyzing the fluctuations of $C_{NI}(.) - C(.)$ consists in splitting it in two summands and in bounding each summand separately:

$$|C(f) - C_{NI}(f)| \leq |C(f) - \mathbb{E}_{p_Z} C_{NI}(f)| + |C_{NI}(f) - \mathbb{E}_{p_Z} C_{NI}(f)|. \tag{13}$$

The first summand bounds the bias induced by taking the smoothed version of the squared error. It is not a stochastic quantity, it should be analyzed using tools from Approximation Theory. This first summand is likely to grow with t. On the contrary, the second term captures the random deviation of C_{NI} with respect to its expectation. Taking advantage that C_{NI} depends smoothly on the sample, it may analyzed using tools from Empirical Process theory (cf. [Ledoux and Talagrand, 1991, chapter 14]). It is expected that this second summand decreases with t.

3.3 Bounding deviations of perturbed cost

As practitioners are likely to minimize $C_{NI^m_{emp}}$ rather than C_{NI}, another difficulty has to be faced: controlling the fluctuations of $C_{NI^m_{emp}}$ with respect to C_{NI}. For a fixed sample, $C_{NI^m_{emp}}$ is a sum of independent random functions

with expectation C_{NI}, in principle it could also be analyzed using Empirical Process techniques.

In the case of sigmoidal neural networks , the boundedness of the summed random variables ensures that for each individual f, $C_{NI^m_{emp}}(f)$ converges almost surely towards $C_{NI}(f)$ as $m \to \infty$, and that $\sqrt{m}\big(C_{NI^m_{emp}}(f) - C_{NI}(f)\big)$ converges in distribution towards a Gaussian random variable. But using Empirical Processes would not pay tribute to the fact that the sampling process defined by NI is under user control, and in our case Gaussian. Sufficiently smooth functions of Gaussian vectors actually obey nice concentration properties as illustrated in the following theorem:

Theorem 3.1 ([Ledoux and Talagrand, 1991]) *Let X be a standard Gaussian vector on $\mathbb{R}^d$. Let f be a Lipschitz function on $\mathbb{R}^d$ with Lipschitz constant smaller than L then:*

$$\mathbb{P}\left\{|f(X) - \mathbb{E}f(X)| > r\right\} \leq 2\exp^{-r^2/2L^2} \quad .$$

Pointwise control of the fluctuations of $C_{NI^m_{emp}}$. If $\mathcal{F}$ is constituted by a class of sigmoidal neural networks, the differentiability assumption for the square loss as a function of inputs is automatically fulfilled.

Assumption 1. In the sequel, we will assume that all weights defining functions in $\mathcal{F}$ are bounded so that: $\mathcal{F}$ is uniformly bounded by some constant M' and the gradient of $f \in \mathcal{F}$ are smaller than some constant L. Let M be a constant greater than $M' + \max_i y^i$.

Then if $\frac{1}{\ell m}\sum_{k=1}^{m}\sum_{i=1}^{\ell}\left(f(\mathbf{x}^i + \boldsymbol{\eta}^{k,i}) - y^i\right)^2$ is regarded as a function on $\mathbb{R}^{\ell m d}$ provided with the Euclidean norm, its Lipschitz constant is upper-bounded by $\frac{2LM}{\sqrt{\ell m}}$.

Applying the preceding theorem to $C_{NI^m_{emp}}(.)$ implies the following observation:

Observation 3.1 *If $\mathcal{F}$ satisfies assumption 1, then:*

$$\mathbb{P}_{\boldsymbol{\eta}}\left\{|C_{NI^m_{emp}}(f) - C_{NI}(f)| > r\right\} \quad \leq 2e^{-m\ell r^2/(8tL^2M^2)}.$$

Remark The dependence of the upper bound on the Lipschitz constant of C_{NI} with respect to $\mathbf{x}$ can not be improved since the theorem 3.1 is tight for linear functions, but the combined dependence on M and L has to be assessed for sigmoidal neural networks. For those networks, large inputs tend to generate small gradients, hence the upper bound provided here may not be tight.

Global control on $C_{NI^m_{emp}}$ Anyhow, pointwise control of $C_{NI^m_{emp}} - C_{NI}$ is insufficient since we need to compare the minimization of $C_{NI^m_{emp}}$ with respect to the minimization of C_{NI}. We may first notice that $\inf_{f\in\mathcal{F}} C_{NI^m_{emp}}(f)$ is a

biased estimator of $\inf_{f\in\mathcal{F}} C_{NI}(f)$:

$$\mathbb{E}_{\boldsymbol{\eta}} \inf_{f\in\mathcal{F}} C_{NI^m_{emp}}(f) \leq \inf_{f\in\mathcal{F}} C_{NI}(f). \tag{14}$$

Second, we may notice that $\inf_{f\in\mathcal{F}} C_{NI^m_{emp}}(f)$ is a concave function of the empirical measure defined by $\boldsymbol{\eta}$, hence it is a backward supermartingale[1], thus it converges almost surely towards a random variable as $m \to \infty$. If $C_{NI^m_{emp}}$ is to converge towards C_{NI}, $\inf_{f\in\mathcal{F}} C_{NI^m_{emp}}(f)$ is due to converge towards $\inf C_{NI}$.

To go beyond this qualitative picture, we need to get a global control on the fluctuations of $\sup_{f\in\mathcal{F}} |C_{NI^m_{emp}}(f) - C_{NI}(f)|$.

Let g denote the function:

$$\boldsymbol{\eta} \mapsto \sup_{f\in\mathcal{F}} |\frac{1}{m\ell} \sum_{i,k} \left(f(\mathbf{x}^i + \boldsymbol{\eta}^{k,i}) - (y^i)\right)^2 - C_{NI}(f)|.$$

If $\mathcal{F}$ satisfies assumption 1, then g is also Lipschitz with coefficient less than $2LM/\sqrt{m\ell}$.

Let us denote $\mathbb{E}_{\boldsymbol{\eta}} \sup_{f\in\mathcal{F}} |C_{NI^m_{emp}}(f) - C_{NI}(f)|$ by E. E may be finite or infinite. E is a function of t, ℓ, m.

Assumption 2. E is finite.

Theorem 3.1 also applies to g , and we get the following concentration result:

Observation 3.2 *If $\mathcal{F}$ is a class of bounded Lipschitz functions satisfying assumption 1 and 2, then*

$$\mathbb{P}_{\boldsymbol{\eta}}\left\{ \sup_{f\in\mathcal{F}} |C_{NI^m_{emp}}(f) - C_{NI}(f)| > E + r \right\} \leq 2e^{-m\ell r^2/(8tL^2M^2)}. \tag{15}$$

This result is obviously only partial in the absence of a nice upper bound on E. It is actually possible to provide explicit upper bounds on E using Metric Entropy techniques and using the subgaussian behavior of the $C_{NI^m_{emp}}$ process described by observation 3.1.

Using the preceding the observation, we may partially control the deviations of approximate minima of $C_{NI^m_{emp}}$ with respect to the infimum of C_{NI}:

Observation 3.3 *If $\mathcal{F}$ satisfies assumptions 1 and 2, then if for any sample $f^\bullet$ satisfies $C_{NI^m_{emp}}(f^\bullet) < \inf_{f\in\mathcal{F}} C_{NI}(f) + \epsilon$ then*

$$\mathbb{E}C_{NI}(f^\bullet) < \inf_{f\in\mathcal{F}} C_{NI}(f) + 2E + \epsilon \ ,$$

and

$$\mathbb{P}_{\boldsymbol{\eta}}\left\{ C_{NI}(f^\bullet) \geq \inf_{f\in\mathcal{F}} C_{NI}(f) + \epsilon + 2(E + r) \right\} \leq 2e^{-m\ell r^2/(8tL^2M^2)}.$$

If ϵ and E can be taken arbitrarily close to 0 using proper tuning of m and t, the corollary shows that minimizing $C_{NI^m_{emp}}$ is a consistent way of approximating the minimum of C_{NI}.

[1]Up to some measurability conditions that are enforced for fixed architecture neural networks.

3.4 Combining global smoothness prior and input perturbation

The derivative-based regularizers proposed in [Bishop, 1995, Leen, 1995] are based on sums of terms evaluated at a finite set of data points. It has been argued that they are not global smoothing priors. At first sight, it may seem that NI escapes this difficulty, anyway, the previous analysis and particularly observation 3.1, as rough as it may be, shows that it may be quite hard to control NI in the absence of any smoothness assumption. Thus it seems cautious to supplement NI or data-dependent derivative-based regularizers with global smoothness constraints.

For sigmoidal neural networks, weight-decay can be combined with NI. The sum M of the absolute values of the weights provides an upper bound on the output values. Let L be the product over the different layers of the sum of absolute values of the weights in those layers, L is an upper bound on the Lipschitz constant of the network. Penalizing by $L + M$, would restrict the search space so that C_{NI} is well-behaved.

Such combinations of penalizers have been reported in the literature [Plaut et al., 1986, Sietsma and Dow, 1991] to be successful. However it seems quite difficult to compare analytically the respective advantages of penalization alone and penalization supplemented with NI. This would require the establishment of lower rate of convergence for one of the techniques. Such rates are not available, up to the authors knowledge.

4 Conclusion and perspective

Under a Gaussian distribution assumption, Noise Injection can be cast as a stochastic alternative to a regularization of the empirical cost by a Partial Differential Equation. This is more likely to facilitate a rigorous analysis of the Noise Injection heuristic than to be a source of efficient algorithms. It allows to assign a precise meaning to the concept of validity of Taylor approximations of the perturbed cost function. For sigmoidal neural networks, and more generally for classes of functions that can be defined using exponentials and polynomials, those Taylor approximation s are indeed valid in the sense that minimizing C_{Taylor} and C_{NI} turn to be equivalent in the infinitesimal variance limit. The practical relevance of this equivalence remains questionable since practical uses of Noise Injection require finite noise variance. The relationship between Noise Injection and the Heat equation enables to establish a simple bridge between regularization and the transformation of C_{emp} into C_{NI}: the contractivity of the Heat semigroup ensures that the regularized version of the effective cost is easier to estimate than the original effective cost. Finally, as practical applications are more likely to minimize empirical versions $C_{NI^m_{emp}}$ of C_{NI} than C_{NI} itself, we resort to results from the theory of Stochastic Calculus to provide bounds on the tail probability of deviations of $C_{NI^m_{emp}}$.

Despite the clarification provided by the "Heat connection", this is far from

being the last word. A lot of quantitative works needs to be done, if theory has to meet practice. The global bounds provided in Section 3.3 need to be complemented by a local analysis focused on the neighborhood of the critical points of C_{emp}. Ultimately, this should provide with rates of convergence for specific $\mathcal{F}$ and classes of target dependencies $\mathbb{E}(Y|\mathbf{x})$. Then as practitioners often use stochastic versions of Back-Propagation, it seems interesting to check whether the approach advocated here can refine the results presented in [An, 1995].

Noise Injection is actually a naive and brutal way of trying to enforce translation invariance while training neural networks. Other (possibly more interesting) forms of invariance under transformation groups deserve to be examined in the Noise Injection perspective, as proposed in [Leen, 1995]. Transformation groups can often be provided with a Riemannian Manifold structure on which some Heat kernels may in turn be defined, thus it is appealing to check whether the generalizations of the tools presented here (the theory of diffusion on manifolds, cf [Ledoux and Talagrand, 1991] for some results) can facilitate the analysis of "Noise Injection" versions of Tangent-Prop like algorithms.

Acknowledgements. We would like to thank two anonymous referees for suggestive comments on an earlier draft of this paper, particularly for suggesting that NI or derivative-based regularizers should be combined with global smoothers. Part of this work has been done while S. Boucheron was visiting the Institute for Scientific Interchange in Torino.

References

[An, 1995] An, G. (1995). The effects of adding noise during backpropagation training on generalization performance. *Neural Computation*, 8:643–674.

[Auer et al., 1996] Peter Auer and Mark Hebster and Manfred K. Warmuth. (1995). Exponentially many local minima for single neurons In *Advances in Neural Information Processing Systems*, 8:316-322.

[Barron et al., 1995] Barron, A., Birge, L., and Massart, P. (1995). Risk bounds for model selection via penalization. Technical report, Universite Paris-Sud. http://www.math.u-psud.fr/stats/preprints.html.

[Bishop, 1995] Bishop, C. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116.

[Comon, 1992] Comon, P. (1992). Classification supervisée par réseaux multicouches. *Traitement du Signal*, 8(6):387–407.

[Devroye, 1987] Devroye, L. (1987). *A Course in Density Estimation*, Birkhäuser, Basel.

[Ethier and Kutrz, 1986] Ethier, S. and Kutrz, T. (1986). *Markov Processes*. J. Wiley and Sons, New York.

[Grandvalet, 1995] Grandvalet, Y. (1995). *Effets de l'injection de bruit sur les perceptrons multicouches.* PhD thesis, Université de Technologie de Compiègne, Compiègne, France.

[Grandvalet and Canu, 1995] Grandvalet, Y. and Canu, S. (1995). A comment on noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(4):678–81.

[Grenander, 1981] Grenander, U. (1981). *Abstract Inference.* J. Wileyand Sons, New York.

[Haussler, 1992] Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and computation*, 100:78–150.

[Hirsch, 1976] Hirsch, M. (1976). *Differential Topology.* Springer Verlag, New York.

[Holmström and Koistinen, 1992] Holmström, L. and Koistinen, P. (1992). Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks*, 3(1):24–38.

[Karatzas and Shreve, 1988] Karatzas, I. and Shreve, S. (1988). *Brownian motion and stochastic calculus*, volume 113 of *GTM.* Springer Verlag, second edition.

[Ledoux and Talagrand, 1991] Ledoux, M. and Talagrand M. (1991). *Probability in Banach Spaces.* Springer Verlag, Berlin.

[Leen, 1995] Leen, T. K. (1995). Data distributions and regularization. *Neural Computation*, 7:974–981.

[Matsuoka, 1992] Matsuoka, K. (1992). Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):436–440.

[Press et al., 1992] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C,* Cambridge University Press, Cambridge.

[Plaut et al., 1986] Plaut, D., Nowlan, S., and Hinton, G. (1986). Experiments on learning by back propagation. Technical Report CMU-CS-86-126, Carnegie-Melon Department of Computer Science, Pittsburgh, PA 15213.

[Reed et al., 1995] Reed, R., Marks II, R., and Oh, S. (1995). Similarities of error regularization, sigmoid gain scaling, target smoothing and training with jitter. *IEEE Transactions on Neural Networks*, 6(3):529–538.

[Sietsma and Dow, 1991] Sietsma, J. and Dow, R. (1991). Creating artificial neural networks that generalize. *Neural Networks*, 4(1):67–79.

[Sontag, 1995] Sontag, E. D. (1995). Critical points for least-squares problems involving certain analytic functions, with applications to sigmoidal neural networks. *Advances in Computational Mathematics* (Special Issue on Neural Networks), to appear.

[Tikhonov and Arsenin, 1977] Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solution of ill-posed problems.* W. H. Wilson, Washington, D. C.

[Vapnik, 1982] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data.* Springer Series in Statistics. Springer-Verlag, New York.

[Webb, 1994] Webb, A. (1994). Functional approximation by feed-forward networks: A least-squares approach to generalization. *IEEE Transactions on Neural Networks*, 5(3):363–371.

From Heat equation to Noise Injection

This appendix rederives the relation between the Heat equation and NI using Fourier analysis techniques. To get the idea in the simplest setting, the problem is treated in one dimension. The Heat equation under concern is:

$$\begin{cases} \dfrac{\partial u}{\partial t} & = & \frac{1}{2}\mathbf{\Delta}_{xx} u \ , \\ u(0,\mathbf{x}) & = & \dfrac{1}{\ell}\displaystyle\sum_{i=1}^{\ell}\left(f(\mathbf{x}^i) - y^i\right)^2 \end{cases}$$

The Fourier transform of the Heat equation is:

$$\frac{\partial \widehat{u}(t,\xi)}{\partial t} = -\ \frac{1}{2}\,\xi^2\ \widehat{u}(t,\xi). \tag{16}$$

This is an ordinary differential equation which solution is:

$$\widehat{u}(t,\xi) = K(\xi)\ \exp\left(-\frac{\xi^2 t}{2}\right) \tag{17}$$

where the integration constant $K(\xi)$ is the Fourier transform of the initial condition, thus:

$$\widehat{u}(t,\xi) = \widehat{u}(0,\xi)\ \exp\left(-\frac{\xi^2 t}{2}\right) \tag{18}$$

The inverse Fourier transform maps the product into a convolution, this entails:

$$u(t,x) = u(0,t) * F^{-1}\left(\exp\left(-\frac{\xi^2 t}{2}\right)\right). \tag{19}$$

Thus, we recover the definition of C_{NI}:

$$C_{NI}(f,t) = C_{emp}(f) * N(t) \tag{20}$$

where $N(t)$ is the density of a centered normal distribution with variance t.

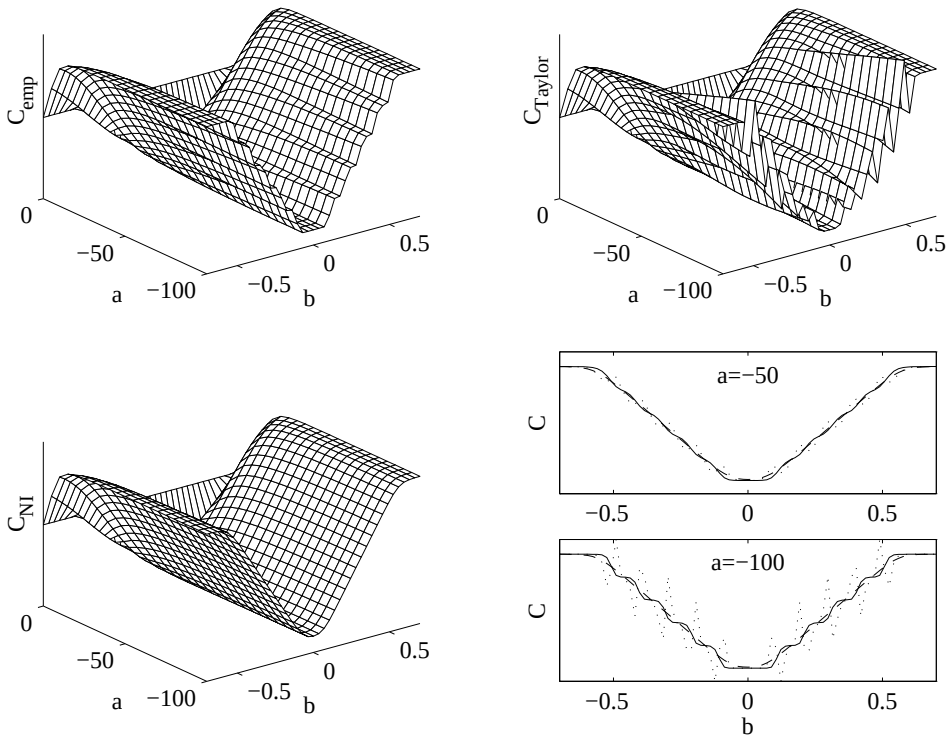

Figure 1: costs C_{emp} (top-left), C_{Taylor} (top-right), and C_{NI} (bottom-left) in the space (a, b). These costs are compared on the bottom-right figures for two values of the parameter a: C_{emp} is plain, C_{Taylor} is dotted, and C_{NI} is dashed.