Communicated by Todd Leen and Christopher Bishop

Noise Injection: Theoretical Prospects

Yves Grandvalet
Stéphane Canu
CNRS UMR 6599 Heudiasyc, Université de Technologie de Compiègne, Compiègne, France

Stéphane Boucheron
CNRS-Université Paris-Sud, 91405 Orsay, France

Noise injection consists of 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: describing analytically the average perturbed cost function is difficult, and controlling the fluctuations of the random perturbed cost function is hard. Hence, recent papers suggest replacing the random perturbed cost by a (deterministic) Taylor approximation of the average perturbed cost function. This article 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 controlling the fluctuations of the random perturbed cost function. Under the 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

Neural network training consists of minimizing a cost functional $C(.)$ on the set of functions $\mathcal{F}$ realizable by multilayer Perceptrons (MLP) with fixed architecture. The cost C is usually the averaged squared error,

$$C(f) = \mathbb{E}_Z \Big(f(X) - Y \Big)^2 \qquad (1.1)$$

where the random variable $Z = (X, Y)$ describing the data is sampled according to a fixed but unknown law. Because C is not computable, an empirically computable cost is then minimized in applications 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}:

$$C_{emp}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mathbf{x}^i) - y^i \right)^2 . \tag{1.2}$$

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 toward 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 of 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 kth epoch of the backpropagation 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:

$$C_{NI^m_{emp}}(f) = \frac{1}{m} \sum_{k=1}^{m} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mathbf{x}^i + \boldsymbol{\eta}^{k,i}) - y^i \right)^2 , \tag{1.3}$$

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

$$C_{NI}(f) = \mathbb{E}_\eta \left[\frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mathbf{x}^i + \boldsymbol{\eta}) - y^i \right)^2 \right] . \tag{1.4}$$

In this article, 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} (see equation 1.4) 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 toward C_{NI}?

Recently, several authors (Webb, 1994; Bishop, 1995; Leen, 1995; Reed, Marks, & Oh, 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}. They not only try to provide a formal description of NI but also aim at finding a deterministic alternative to the minimization of $C_{NI^m_{emp}}$.

This article 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 dependence of C_{NI} (see equation 1.4) on the noise variance is shown to obey the heat equation (see section 2.1). This clear connection between partial differential equations and NI provides some indications on the relevance domain of traditional Taylor expansions (see section 2.2). Finally, we analyze the simplified expressions that are assumed to be valid locally around optimal solutions (see section 2.3).

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

2 Taylor Expansions

2.1 Gaussian Perturbation and Heat Equation. To exhibit the connection between gaussian NI and the heat equation, let us define u as a function from $\mathbb{R}^+ \times \mathbb{R}^{\ell \times d}$ by:

$$u(0,\mathbf{x}) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(f(\mathbf{x}^i) - y^i\right)^2 \tag{2.1}$$

$$u(t,\mathbf{x}) = \mathbb{E}_{\boldsymbol{\eta}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}\left(f(\mathbf{x}^i + \boldsymbol{\eta}^i) - y^i\right)^2\right]. \tag{2.2}$$

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 $\left(T_t\right)_{t\geq 0}$ defines a semigroup, the heat semigroup (cf. Ethier & Kutrz, 1986, for an introduction to semigroup operators). The function u obeys the heat equation (cf. Karatzas & Shreve, 1988, chap. 4, sec. 3, 4):

$$\frac{\partial u}{\partial t} = \frac{1}{2}\boldsymbol{\Delta}_{\mathbf{xx}}u \tag{2.3}$$

where $\mathbf{\Delta_{xx}}$ is the Laplacian with respect to $\mathbf{x}$, and where the initial conditions are defined by equation 2.1. For the sake of self-containment, a derivation of equation 2.3 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; equation 2.3 yields:

$$C_{NI}(f, t) = C_{emp}(f) + \frac{1}{2}\int_0^t \mathbf{\Delta_{xx}} C_{NI}(f, s)\, ds. \tag{2.4}$$

Therefore, C_{NI} can be investigated in the purely analytical framework of partial differential equations (under the gaussian assumption). The possibility of forgetting about the original probabilistic setting when dealing with neural networks follows from the Tikhonov uniqueness theorem (Karatzas & Shreve, 1988, chap. 4, sec. 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 (2.1) with initial conditions defined in equation 2.1.

Any deterministic faithful simulation of NI should use some numerical analysis software to integrate the heat equation and then run some backpropagation software on the result. We do not recommend such a methodology for efficiency reasons and insist that stochastic representations of solutions of partial differential equations (PDEs) have proved useful in analysis (Karatzas & Shreve, 1988).

Methods reported in the literature (such as finite differences and finite element; cf. Press, Teukolsky, Vetterling, & Flannery, 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 the 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), and Reed et al. (1995) that:

$$C_{Taylor}(f) = C_{emp}(f) + \frac{\sigma^2}{2}\, \mathbf{\Delta_{xx}} C_{emp}(f). \tag{2.5}$$

In the context of gaussian NI, this means that the Laplacian is the infinitesimal generator of the heat semigroup. To emphasize the distinction between equations 2.5 and 2.3, one should stress the fact that the heat equation is not only a correct description of the impact of gaussian NI 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

$$\lim_{\sigma^2 \to 0} \frac{C_{NI}(f) - C_{Taylor}(f)}{\sigma^2} = 0. \quad (2.6)$$

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 & Shreve, 1988, theorems 3.3, 4.2, chap. 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) and Leen (1995). The first issue cannot 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 toward the minima of C_{emp} when $t = \sigma^2 \to 0$. This cannot be deduced directly from equation 2.6 since it describes only simple convergence. A uniform convergence over $\mathcal{F}$ is required. As we will vary t, let us denote by $C_{NI}(f, t)$ (respectively $C_{Taylor}(f, t)$) the perturbed cost (resp. its Taylor approximation) of f when the noise variance is t, we get:

$$C_{NI}(f, t) - C_{Taylor}(f, t) = \frac{1}{2} \int_0^t \left[\boldsymbol{\Delta}_{\mathbf{xx}} C_{NI}(f, s) - \boldsymbol{\Delta}_{\mathbf{xx}} C_{emp}(f) \right] ds. \quad (2.7)$$

If some uniform (over $\mathcal{F}$ and $s \le t_0 < 0$) bound on $|\boldsymbol{\Delta}_{\mathbf{xx}} C_{NI}(f, s) - \boldsymbol{\Delta}_{\mathbf{xx}} C_{emp}(f)|$ is available, then $\lim_{t \to 0} \max_{f \in \mathcal{F}} C_{NI}(f, t) - C_{Taylor}(f, t) = 0$, and small manipulations using the triangular inequality reveal the convergence of minima. The same argument shows the convergence of the minima of C_{NI} toward 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 ensure 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

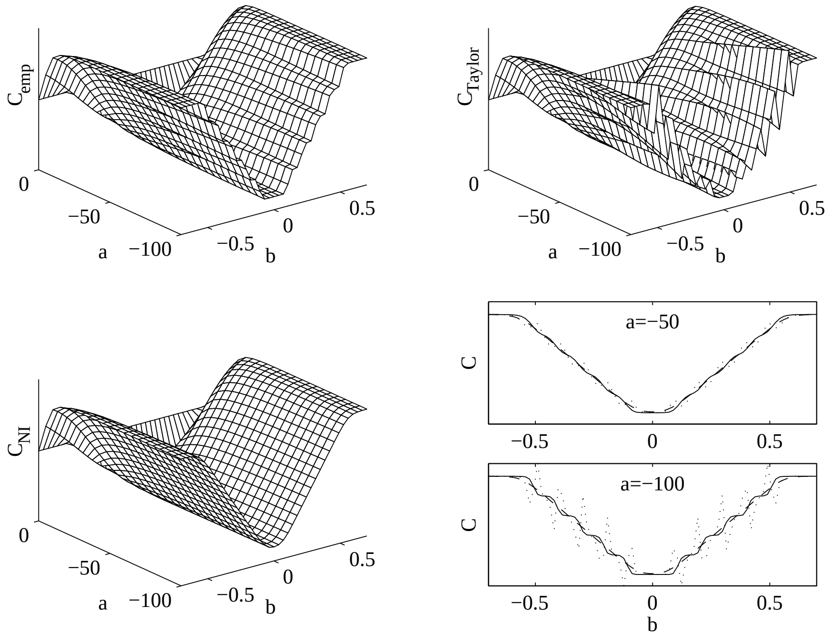

Figure 1: Costs C_{emp} (top left), C_{Taylor} (top right), and C_{NI} (bottom left) in the space (a, b). These costs are compared on the bottom-right figures for two values of the parameter a: C_{emp} is solid, C_{Taylor} is dotted, and C_{NI} is dashed.

$f(x + \eta) = g(wx + w\eta)$. 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 nonlinear (i.e., f nonlinear), the Taylor expansion will fail for large w^2.

A simple illustration is given in Figure 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 $f(x) = \exp(ax + b)\big/(1 + \exp(ax + b))$ $a, b \in \mathbb{R}$, 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 nonzero 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 points in the sample.

Remark. Although equation 2.5 has been established for various kinds of noise (strong integrability properties are sufficient), C_{NI} is the solution of the heat equation only for gaussian noise. In nongaussian contexts NI usually does not define a semigroup of operators indexed by variance.

2.3 Local Analysis and Simplified Expressions. Requiring uniform bounds on $|\mathbf{\Delta}_{\mathbf{xx}}C_{NI}(f,s) - \mathbf{\Delta}_{\mathbf{xx}}C_{emp}(f)|$ may be too demanding. Fortunately, it is possible to adapt the local approach to NI suggested in Leen (1995) and 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}. Although Bishop and Leen focused their attention on the case where $\mathbb{E}(Y|X = \mathbf{x})$ is realizable by the neural architecture and the empirical measure closely mimics the sampling law (e.g., Leen, 1995, sec. 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 nondegenerate iff the Hessian matrix has full rank. In the sequel, we will assume that C_{emp} has only nondegenerate 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 nondegenerate critical points is finite (Sontag, 1996).

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 (1995) suggestion. 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 nondegenerate 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, p. 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 norms 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. 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 nondegenerate minima, which are in one-to-one correspondence and such that the convergence of the minima of C_{NI} toward the minima of C_{emp} as t tends toward 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 an explanation resort to the concepts of penalization and regularization. Usually penalization consists of adding a positive functional $\Omega(.)$ to the empirical risk. It is called a regularization if the sets $\{f\colon\ f \in \mathcal{F},\quad \Omega(f) \leq \alpha\}$ are compact for the topology on $\mathcal{F}$. Penalization and regularization are standard ways of improving regression and estimation techniques (e.g., Grenander, 1981; Vapnik, 1982; Barron, Birge, & Massart, 1995).

When $\mathcal{F}$ is the set of linear functions, NI has been recognized as a regularizer in the Tikhonov sense (Tikhonov & Arsenin, 1977). This is the only reported case, but it is enough to consider $\mathcal{F}$ as constituted by univariate degree 2 polynomial to realize that NI cannot 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 NI with kernel estimation techniques (Comon, 1992; Holmström & 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, Hebster, & Warmuth, 1996). Hence the safest (and, to our 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 $C(.)$, that is, on $\sup_{\mathcal{F}} |C_{NI}(f) - C(f)|$. The poor performance of minimization of empirical risk could be due to the slow convergence of the empirical measure $\hat{p}_Z \overset{\Delta}{=} \sum_i \delta_{\mathbf{x}^i, y^i}$ toward the sampling probability in the pseudometric induced by $\mathcal{F}$ $(d_{\mathcal{F}}(\hat{p} - \hat{p}') \overset{\Delta}{=} \sup_{f \in \mathcal{F}} |\mathbb{E}_{\hat{p}}(f(X) - Y)^2 - \mathbb{E}_{\hat{p}'}(f(X) - Y)^2|)$.

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

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \mathbb{E}_{\boldsymbol{\eta}, \boldsymbol{\eta}'} \left(f(\mathbf{x}^i + \boldsymbol{\eta}^i) - (y^i + \boldsymbol{\eta}') \right)^2, \tag{3.1}$$

where $\boldsymbol{\eta}'$ is a scalar gaussian independent of $\boldsymbol{\eta}$. Minimizing C_{NI} consists of minimizing the empirical risk against a smoothed version of the empirical measure: the Parzen-Rosenblatt estimator of the density (for details on the latter, see Devroye, 1987). The connection with gaussian kernel density estimation and regularization can thus be established in a perspective described in Grenander (1981). Because the gaussian kernel is the fundamental solution of the heat equation, the smoothed density that defines equation 3.1 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 solving 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 of splitting it in two summands and bounding each summand separately:

$$|C(f) - C_{NI}(f)| \leq |C(f) - \mathbb{E}_{p_Z} C_{NI}(f)| + |C_{NI}(f) - \mathbb{E}_{p_Z} C_{NI}(f)|. \qquad (3.2)$$

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 be analyzed using tools from empirical process theory (cf. Ledoux & Talagrand, 1991, chap. 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 toward $C_{NI}(f)$ as $m \to \infty$ and that $\sqrt{m}\big(C_{NI^m_{emp}}(f) - C_{NI}(f)\big)$ converges in distribution toward 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 (Tsirelson, 1976; c. f. Ledoux & 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:*

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

3.3.1 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 $f \in \mathcal{F}$ is smaller than some constant L. Let M be a constant greater than $M' + \max_i y^i$.

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

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

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

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

Remark. The dependence of the upper bound on the Lipschitz constant of C_{NI} with respect to **x** cannot be improved since 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.

3.3.2 Global Control on $C_{NI^m_{emp}}$. 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 $\inf_{f\in\mathcal{F}} C_{NI}(f)$:

$$\mathbb{E}_{\boldsymbol{\eta}} \inf_{f\in\mathcal{F}} C_{NI^m_{emp}}(f) \leq \inf_{f\in\mathcal{F}} C_{NI}(f). \tag{3.3}$$

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] and thus converges almost surely toward a random variable as $m \to \infty$. If $C_{NI^m_{emp}}$ is to converge toward C_{NI}, $\inf_{f\in\mathcal{F}} C_{NI^m_{emp}}(f)$ is due to converge toward $\inf C_{NI}$.

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:

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

If $\mathcal{F}$ satisfies assumption 1, then g is also Lipschitz with a 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.

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

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 assumptions 1 and 2, then

$$\mathbb{P}_{\boldsymbol{\eta}}\left\{\sup_{f\in\mathcal{F}} |C_{NI^m_{emp}}(f) - C_{NI}(f)| > E + r\right\} \leq 2e^{-m\ell r^2/(8tL^2M^2)}. \tag{3.4}$$

This result is only partial in the absence of a good upper bound on E. It is 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 observation 3.2, 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}.

3.4 Combining Global Smoothness Prior and Input Perturbation.

The derivative-based regularizers proposed in Bishop (1995) and 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 be that NI escapes this difficulty; 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, Nowlan, & Hinton, 1986; Sietsma & 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 establishing lower rate of convergence for one of the techniques. Such rates are not available to our knowledge.

4 Conclusion and Perspective

Under a gaussian distribution assumption, NI 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 NI heuristic than to be a source of efficient algorithms. It allows the assignment of 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 approximations are valid in the sense that minimizing C_{Taylor} and C_{NI} turns out to be equivalent in the infinitesimal variance limit. The practical relevance of this equivalence remains questionable since practical uses of NI require finite noise variance. The relationship between NI and the heat equation enables establishing 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, because 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 is 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 rates of convergence for specific $\mathcal{F}$ and classes of target dependencies $\mathbb{E}(Y|\mathbf{x})$. Because practitioners often use stochastic versions of backpropagation, it will be interesting to see whether the approach advocated here can refine the results presented in An (1995).

NI is a naive and rather conservative way of trying to enforce translation invariance while training neural networks. Other, and possibly more interesting, forms of invariance under transformation groups deserve to be examined in the NI perspective, as proposed by Leen (1995). Transformation groups can often be provided with a Riemannian manifold structure on which some heat kernels may be defined; thus it is appealing to check whether the generalizations of the tools presented here (the theory of diffusion on manifolds; cf. Ledoux & Talagrand, 1991, for some results) can facilitate the analysis of NI versions of tangent-prop–like algorithms.

Appendix: From Heat Equation to Noise Injection

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

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

The Fourier transform of the heat equation is:

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

This is an ordinary differential equation whose solution is:

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

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

$$\widehat{u}(t,\xi) = \widehat{u}(0,\xi)\ \exp\left(-\frac{\xi^2 t}{2}\right). \tag{A.3}$$

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

$$u(t,x) = u(0,t) * F^{-1}\left(\exp\left(-\frac{\xi^2 t}{2}\right)\right). \tag{A.4}$$

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

$$C_{NI}(f,t) = C_{emp}(f) * N(t), \tag{A.5}$$

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

Acknowledgments

We thank two anonymous referees for helpful comments on a draft of this article, particularly for suggesting that NI or derivative-based regularizers should be combined with global smoothers. Part of this work has been done while Stéphane Boucheron was visiting the Institute for Scientific Interchange in Torino. Boucheron has been partially supported by the German-French PROCOPE program 96052.

References

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

Auer, P., Hebster, M., & Warmuth, M. K. (1996). Exponentially many local minima for single neurons. *Advances in Neural Information Processing Systems 8*:316–322.

Barron, A., Birge, L., & Massart, P. (1995). *Risk bounds for model selection via penalization* (Tech. Report). Université Paris-Sud. http://www.math.u-psud.fr/stats/preprints.html.

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

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

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

Ethier, S., & Kutrz, T. (1986). *Markov Processes*. New York: Wiley.

Grandvalet, Y. (1995). *Effets de l'injection de bruit sur les perceptrons multicouches*. Unpublished doctoral dissertation, Université de Technologie de Compiègne, Compiègne, France.

Grandvalet, Y. & Canu, S. (1995). A comment on noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics 25*(4):678–681.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

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

Hirsch, M. (1976). *Differential Topology*. New York: Springer-Verlag.

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

Karatzas, I., & Shreve, S. (1988). *Brownian Motion and Stochastic Calculus* (2nd ed.). New York: Springer-Verlag.

Ledoux, M., & Talagrand, M. (1991). *Probability in Banach Spaces*. Berlin: Springer-Verlag.

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

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

Plaut, D., Nowlan, S., & Hinton, G. (1986). *Experiments on learning by back propagation* (Tech. Rep. CMU-CS-86-126). Pittsburgh, PA: Carnegie Mellon University, Department of Computer Science.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.

Reed, R., Marks II, R., & 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, J., & Dow, R. (1991). Creating artificial neural networks that generalize. *Neural Networks 4*(1):67–79.

Sontag, E. D. (1996). Critical points for least-squares problems involving certain analytic functions, with applications to sigmoidal neural networks. *Advances in Computational Mathematics 5*:245–268.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of Ill-Posed Problems*. Washington, DC: W. H. Wilson.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag.

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

Received February 1, 1996; accepted September 4, 1996.