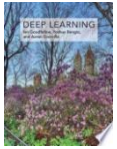
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Implicit regularization in machine learning: a perspective from iterative regularization theory

Martin Burger, FAU Erlangen-Nürnberg

Learning and Regularization



Chapter 7

Regularization for Deep Learning

Various aspects of regularization in machine learning

Explicit regularization (variational, dropout etc) or implicit regularization via (stochastic) gradient descent

Theoretical understanding ?

Relation to regularization theory for ill-posed problems ?

Basic setup

Consider parametrized model, e.g. deep neural network

$$y = f(x; \theta)$$

Parameter determined by (approximate) minimization of empirical risk

$$E(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i; \theta), y_i)$$

Theoretical understanding ?

Relation to regularization theory for ill-posed problems ?

Empirical and Population Risk

Empirical risk $$E(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i; \theta), y_i)$$

Standard assumption: i.i.d. sampling of $z_i = (x_i, y_i)$

Underlying probability measure *P* used for sampling

Population risk

$$F(\theta) = \mathbb{E}_P(\ell(f(x; \theta), y))$$

Generalization error

$$G(\theta) = F(\theta) - E(\theta)$$

Empirical risk minimization and generalization

Statistical analysis of empirical risk minimization:

Estimate of the population risk or generalization error
(in expectation / distribution / confidence)
in terms of N …

and other parameters (network width, layer size, VC dimension …)

Empirical risk minimization and generalization

A Priori Estimates of the Population Risk for Residual Networks

Weinan E[1,2,3], Chao Ma[2], and Qingcan Wang[2]

Theorem 2.5 (A priori estimate). *Let $f^* : \Omega \to [0,1]$ and assume that the residual network $f(\cdot;\boldsymbol{\theta})$ has architecture (2.1). Let n be the number of training samples, L be the number of layers and m be the width of the residual blocks. Let $\mathcal{L}(\boldsymbol{\theta})$ and $\hat{\mathcal{L}}(\boldsymbol{\theta})$ be the truncated population risk and empirical risk defined in (2.4) respectively; let $\|f\|_{\mathcal{B}}$ be the Barron norm of f^* and $\|\boldsymbol{\theta}\|_{\mathrm{P}}$ be the weighted path norm of $f(\cdot;\boldsymbol{\theta})$ in Definition 2.1 and 2.4. For any $\lambda \geq 4 + 2/[3\sqrt{2\log(2d)}]$, assume that $\hat{\boldsymbol{\theta}}$ is an optimal solution of the regularized model*

$$\min_{\boldsymbol{\theta}} \ \mathcal{J}(\boldsymbol{\theta}) := \hat{\mathcal{L}}(\boldsymbol{\theta}) + 3\lambda\|\boldsymbol{\theta}\|_{\mathrm{P}}\sqrt{\frac{2\log(2d)}{n}}. \tag{2.12}$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random training samples, the population risk satisfies

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) \leq \frac{16\|f\|_{\mathcal{B}}^2}{Lm} + (12\|f\|_{\mathcal{B}} + 1)\frac{3(4+\lambda)\sqrt{2\log(2d)} + 2}{\sqrt{n}} + 4\sqrt{\frac{2\log(14/\delta)}{n}}. \tag{2.13}$$

Empirical risk minimization and generalization

An Optimal Transport View on Generalization

Jingwei Zhang* Tongliang Liu* Dacheng Tao*

Theorem 6. *If $\mathcal{W}$ has finite VC-dimension d, the expected generalization error of a learning algorithm $\mathcal{A}$ for binary classification can be upper bounded by*

$$\mathbb{E}[R(W) - R_{S_n}(W)] \leq 2 * \mathbb{TV}(P_W \times P_z, P_{W,z}) \leq \sqrt{\frac{2d \log_+(\frac{ne}{d})}{n}}\,, \tag{42}$$

where $\log_+(x) := \max\{1, \log x\}$.

Empirical risk minimization and generalization

Train faster, generalize better:
Stability of stochastic gradient descent

Moritz Hardt* Benjamin Recht† Yoram Singer‡

Theorem 2.2. *[Generalization in expectation] Let A be ϵ-uniformly stable. Then,*

$$|\mathbb{E}_{S,A}\left[R_S[A(S)] - R[A(S)]\right]| \leq \epsilon\,.$$

Theorem 3.12. *Assume that $f(\cdot; z) \in [0,1]$ is an L-Lipschitz and β-smooth loss function for every z. Suppose that we run SGM for T steps with monotonically non-increasing step sizes $\alpha_t \leq c/t$. Then, SGM has uniform stability with*

$$\epsilon_{\text{stab}} \leq \frac{1 + 1/\beta c}{n-1}(2cL^2)^{\frac{1}{\beta c+1}} T^{\frac{\beta c}{\beta c+1}}$$

Regularization Theory

What has this to do with regularization theory ?

Basic ideas and assumptions of regularization theory:

- Existence of an ***ideal solution***
- Application of regularization method to ***noisy data***
- Convergence of regularized solution / error estimate between regularized and ideal solution

Ideal solution as minimizer of ideal likelihood

$$D(\theta) = \mathbb{L}(A(\theta), b)$$

Regularization Theory

Example: variational regularization of inverse problems

Regularized solution as minimizer of

$$J_\alpha(\theta) = D'(\theta) + \alpha R(\theta)$$

Perturbation in *D* due to errors in *A* and *b*

Example: regularization by gradient descent

$$\partial_t \theta = -\nabla_\theta D'(\theta)$$

Regularization from appropriate early stopping at time T depending on the noise strength

Inverse Problems and Empirical Risk Minimization

Standard formulations of inverse problems can be recast as (regularized) empirical risk minimization

Example: Radon transform in 2D, standard parametrization of lines by angle and distance to origin $x \in [0, \pi) \times [0, L]$

Take uniform distribution for line parameter, Gaussian additive noise at each point. Empirical risk equals standard likelihood (plus constant)

$$F(\theta) = \frac{1}{2L\pi} \int_{[0,\pi)\times[0,L]} \int_{\mathbb{R}} |\mathcal{R}(\theta)(x) - \mathcal{R}(\theta^*)(x) - n|^2 \, dR(n) \, dx$$

$$= \frac{1}{2L\pi} \int_{[0,\pi)\times[0,L]} |\mathcal{R}(\theta)(x) - \mathcal{R}(\theta^*)(x)|^2 \, dx + \int_{\mathbb{R}} n^2 \, dR(n)$$

Error formulation for risk minimization

Idea: measurement error is an error in the distribution *P*

Generalized definition of empirical risk

$$E(\theta) = \mathbb{E}_{P'}(\ell(f(x;\theta), y))$$

Perturbation of ideal distribution by sampling, but also adversarial examples, measurement bias, …

Noise level = distance between probability measures (distributional robustness)

What is a canonical distance ?

Sources of Perturbed Distributions

- Sampled Measurements

$$P' = \frac{1}{N}\sum_{i=1}^{N}\delta_{z_i}$$

Distance measure needs to be able to deal with empirical measures

Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance

Jonathan Weed*,† and Francis Bach†

Massachusetts Institute of Technology
INRIA – ENS

ON THE RATE OF CONVERGENCE IN WASSERSTEIN DISTANCE OF THE EMPIRICAL MEASURE

NICOLAS FOURNIER AND ARNAUD GUILLIN

Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces

Jing Lei[1]

Various estimates available in Wasserstein metrics

$$W_p(P,P') = \left(\inf_{\Pi\in\Gamma(P,P')} \mathbb{E}_{(z,z')\sim\Pi}\left(\|z-z'\|^p\right)\right)^{1/p}$$

Sources of Perturbed Distributions

- Adversarial examples

Small movement of existing samples, naturally bounded in Wasserstein distances

- Phantom measurements in imaging

Input data definitely not from the correct distributions of images

The use of Wasserstein metrics

Natural assumption: Lipschitz continuity of map depending on parameters

$$\|f(x;\theta) - f(x';\theta)\| \leq L_f(\theta)\ \|x - x'\|$$

Basic relation to optimal transport

$$G(\theta) = \mathbb{E}_{(z,z')\sim\Pi}\left(\ell(f(x;\theta),y) - \ell(f(x';\theta),y')\right)$$

for each transport plan

Wasserstein estimate: Lipschitz loss

With straight-forward proof

$$G(\theta) \leq L_\ell \ \max\{L_f(\theta), 1\} W_1(P, P') \leq L_\ell \ \max\{L_f(\theta), 1\} W_p(P, P')$$

Estimate holds for arbitrary parameter, not just minimizers

Standard finding: small Lipschitz constant improves generalization

Bounds could be gained using other transport distances more closely related to the architecture

Wasserstein estimate: Square loss

Introduce a generalized generalization error

$$G_\epsilon(\theta) = F(\theta) - (1+\epsilon)E(\theta)$$

Suitable in overparametrized settings, when *E* is very small

$$Let\ \ell = \tfrac{1}{2}\|\cdot\|^2,\ then\ for\ each\ \epsilon > 0$$

$$G_\epsilon(\theta) \leq \frac{2+\epsilon}{4\epsilon}\ \max\{L_f(\theta)^2, 1\}\ W_2(P,P')^2 \leq \frac{2+\epsilon}{4\epsilon}\max\{L_f(\theta)^2, 1\}\ W_p(P,P')^2$$

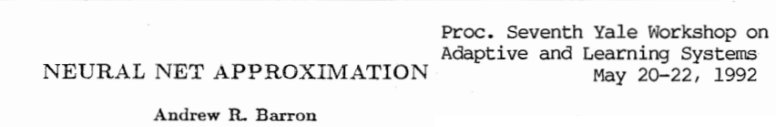
What about ill-posedness ?

Problem of estimating parameters is highly ill-posed

Classical example: two-layer network with stochastic inner weights

$$f(x;\theta) = \frac{1}{N}\sum_{i=1}^{N} \theta_i^2 \sigma(\theta_i^1 \cdot x + \theta_i^0)$$

Can be interpreted as random discretization of first kind integral operator


NEURAL NET APPROXIMATION

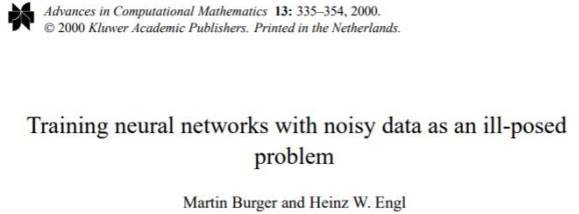
Andrew R. Barron

Proc. Seventh Yale Workshop on
Adaptive and Learning Systems
May 20-22, 1992

$$F(x) = \int \theta^2(a,b)\sigma(a \cdot x + b)\ d\mu(a,b)$$

What does generalization error estimate ?

For large *N* even the problem of estimating outer weights is asymptotically highly ill-posed


Advances in Computational Mathematics **13:** 335–354, 2000.

Training neural networks with noisy data as an ill-posed problem

Martin Burger and Heinz W. Engl

Additional nonlinear ill-posedness even in finite dimensions for inner weights (mb-Haslinger-Bodenhofer-Engl 02)

In most examples of learning the **parameters are of no relevance**
Only the output is interesting, approximation of the map *F*

Estimating *F* in a norm related to empirical risk is well-posed !
Mildly ill-posed in stronger norm

What does generalization error estimate ?

Consider ideal solution

$$F^* = \arg\min_{F \in \mathcal{F}} \mathbb{E}_{(x,y)\sim P}(\ell(F(x), y))$$

Optimality condition for all *F*

$$\mathbb{E}_P(\partial_F \ell(F^*(x), y)(F(x) - F^*(x)) = 0$$

Modern regularization theory estimates error in Bregman distance

$$d(f, F^*) = \mathbb{E}_P(\ell(f(x;\theta), y) - \ell(F^*(x), y) - \partial_F \ell(F^*(x), y)(f(x;\theta) - F^*(x)))$$

What does generalization error estimate ?

Insert optimality and decompose Bregman distance

$$\begin{aligned} d(f, F^*) =& \mathbb{E}_P(\ell(f(x;\theta), y)) - \mathbb{E}_{P'}(\ell(f(x;\theta), y)) + \\ & \mathbb{E}_{P'}(\ell(f(x;\theta), y) - \ell(F^*(x), y)) + \\ & \mathbb{E}_{P'}(\ell(F^*(x), y)) - \mathbb{E}_P(\ell(F^*(x), y)) \end{aligned}$$

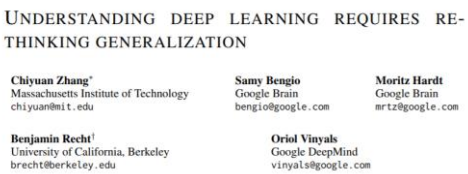
Sum of generalization error, approximation error, noise or sampling error

Approximation error only on training set, potentially negative
Sampling error independent of training result

Generalization error measures effect of training quality on distance

Implicit regularization

Observation in practice: solution obtained with (stochastic) gradient descent not iterated until convergence have good generalization properties



UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

IN SEARCH OF THE REAL INDUCTIVE BIAS: ON THE ROLE OF IMPLICIT REGULARIZATION IN DEEP LEARNING

Behnam Neyshabur, Ryota Tomioka & Nathan Srebro
Toyota Technological Institute at Chicago
Chicago, IL 60637, USA
{bneyshabur,tomioka,nati}@ttic.edu

Generalization Properties and Implicit Regularization for Multiple Passes SGM

Junhong Lin*
Raffaello Camoriano†,*,‡
Lorenzo Rosasco*,‡
JHLIN5@HOTMAIL.COM
RAFFAELLO.CAMORIANO@IIT.IT
LROSASCO@MIT.EDU
*LCSL, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia, Cambridge, MA 02139, USA
‡DIBRIS, Università degli Studi di Genova, Via Dodecaneso 35, Genova, Italy
†iCub Facility, Istituto Italiano di Tecnologia, Via Morego 30, Genova, Italy

Relation to iterative regularization methods for ill-posed problems: iterates decrease distance to exact solution until residual is too small (discrepancy principle)