

SMOOTHNESS AND EXTREME VALUES OF ERROR SURFACES

Michael Turmon

7 December 1996

- A. Introduction
- B. Error Surface as a Stochastic Process
- C. Poisson Clumping: Method
- D. Poisson Clumping: Results
- E. Conclusions

(Work performed partly at the Jet Propulsion Laboratory,
California Institute of Technology, under contract with the
National Aeronautics and Space Administration.)

mjt@aig.jpl.nasa.gov

<http://www-aig.jpl.nasa.gov/home/mjt/>

PROBLEM SETTING

- Nets $\eta(x; w)$ take vector inputs $x \in R^p$ to $\{0, 1\}$ or R .
Weight vector $w \in \mathcal{W} \subseteq R^d$.
Training set $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from P .

- Using (the unknown) P , define mean-squared error

$$\begin{aligned}\mathcal{E}(w) &= E(\eta(x; w) - y)^2 \\ &= P(\eta(x; w) \neq y) \quad (\text{for classifiers})\end{aligned}$$

- Choose classifier via the (accessible) error surface

$$\nu_{\mathcal{T}}(w) = \frac{1}{n} \sum_{i=1}^n (\eta(x_i; w) - y_i)^2$$

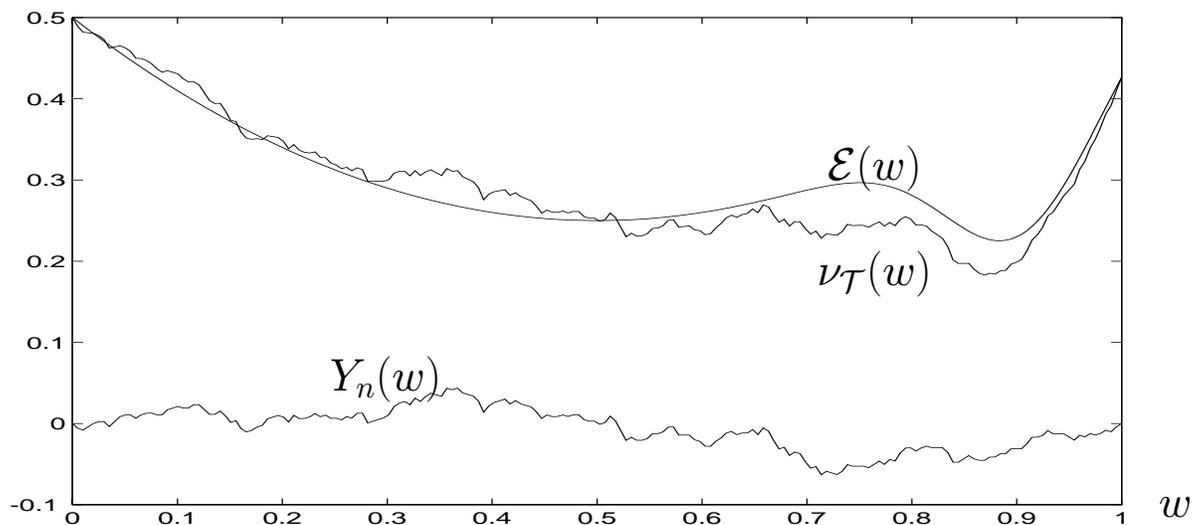
- We choose to examine disagreement

$$Y_n(w) = \nu_{\mathcal{T}}(w) - \mathcal{E}(w)$$

$$\nu_{\mathcal{T}}(w) = \mathcal{E}(w) + Y_n(w)$$

Especially concerned about large disagreements

PROPERTIES OF DISAGREEMENT



If $\mathcal{N} = \{\eta(\cdot; w)\}$ is a VC class, have uniform SLLN

$$\sup_{w \in \mathcal{W}} |Y_n(w)| \rightarrow 0 \quad \text{a.s.}$$

Significance of $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$

- For generalization: find n such that

$$\sup_{w \in \mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \leq \epsilon \quad \text{w.p. } \tau \text{ near } 1$$

- Roughness and extrema of discrepancy influence training

Two related questions

- How fast does $Y_n(w) \rightarrow 0$?
- What does $Y_n(w)$ look like in weight space?

RESCALING THE DISCREPANCY

Let $\sigma^2(w) = \text{Var}((y - \eta(x; w))^2)$.

Classification: $\sigma^2(w) = \mathcal{E}(w)(1 - \mathcal{E}(w))$.

Consider as a function of w

$$\frac{\nu_{\mathcal{T}}(w) - \mathcal{E}(w)}{\sigma(w)} \quad \text{vs.} \quad \nu_{\mathcal{T}}(w) - \mathcal{E}(w)$$

- Large-variance weights dominate the sample path
Undesirable: such nets are bad models
E.g. classification maximum variance at $\mathcal{E}(w) = 1/2$
- Normalization simultaneously provides
greater resolution around $\mathcal{E}(w) = 0$.

Suppose with high probability, in classification,

$$\sup_w \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sigma(w)} \leq \epsilon$$

Then if $\nu_{\mathcal{T}}(w) = 0$,

$$\mathcal{E}(w) \leq \epsilon^2 / (1 + \epsilon^2) < \epsilon^2$$

The condition tightens considerably

PRIOR WORK

Vapnik Bound

Vapnik:

$$P(\| |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \|_{\mathcal{W}} > \epsilon) \leq 6 \left(\frac{2en}{v} \right)^v e^{-n\epsilon^2/4}$$

- no unknown constant factors
- independent of P
- independent of y

Discrepancy shrinks as

$$\epsilon = \left(4 \frac{v}{n} \log \frac{2en}{v} \right)^{1/2}$$

Empirical Processes

Talagrand:

$$P(\| |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \|_{\mathcal{W}} > \epsilon) \leq K_1 \left(\frac{K_2 n \epsilon^2}{v} \right)^v e^{-2n\epsilon^2}$$

provided $n \geq K_3 v / \epsilon^2$ (for any VC class)

Discrepancy shrinks as

$$\epsilon = \left(\frac{v}{n} (K_3 \vee (1/2) \log K_2) \right)^{1/2}$$

but result is of no immediate use

The order of dependence is $(v/n)^{1/2}$, with no log factor.

FURTHER DEVELOPMENTS

Normalized Error

Vapnik:

$$P(\|(\mathcal{E}(w) - \nu_{\mathcal{T}}(w))/\sqrt{\mathcal{E}(w)}\|_{\mathcal{W}} > \epsilon) \leq 8\left(\frac{2en}{v}\right)^v e^{-n\epsilon^2/4}$$

Scaled discrepancy shrinks as

$$\epsilon = \left(4 \frac{v}{n} \log \frac{2en}{v}\right)^{1/2}$$

above which with high probability

$$(\forall w \in \mathcal{W}) \frac{\mathcal{E}(w) - \nu_{\mathcal{T}}(w)}{\sqrt{\mathcal{E}(w)}} \leq \epsilon$$

Prediction of an $O((v/n) \log(n/v))$ shrinkage when working near $\mathcal{E}(w) = 0$.

Another Approach

View $\mathcal{E}(w) - \nu_{\mathcal{T}}(w)$ as a random process indexed by $w \in \mathcal{W}$.

Use the smoothness properties of that process to characterize extremes of the discrepancy

NORMAL APPROXIMATION

Application of CLT

- For large n the CLT tells us

$$Z_n(w) := \sqrt{n} Y_n(w) = \sqrt{n} [\nu_{\mathcal{T}}(w) - \mathcal{E}(w)]$$

is nearly Gaussian.

- Functional CLT says the same for $\{Z_n(w)\}$, $w \in \mathcal{W}$.
- Not tail behavior (probabilities not vanishingly small)

Normal Process

$$\sqrt{n} [\nu_{\mathcal{T}}(w) - \mathcal{E}(w)] \stackrel{\mathcal{D}}{\approx} Z(w)$$

$Z(w)$ is the mean-zero normal process defined by

$$R(w, v) = E Z(w)Z(v) = \text{Cov}((y - \eta(x; w))^2, (y - \eta(x; v))^2)$$

Also define

$$\sigma^2(w) = R(w, w)$$

$$\bar{\Phi}(b) = P(N(0, 1) > b)$$

Summary

Problem of extrema of the empirical process corresponds to one about extrema of a Gaussian process

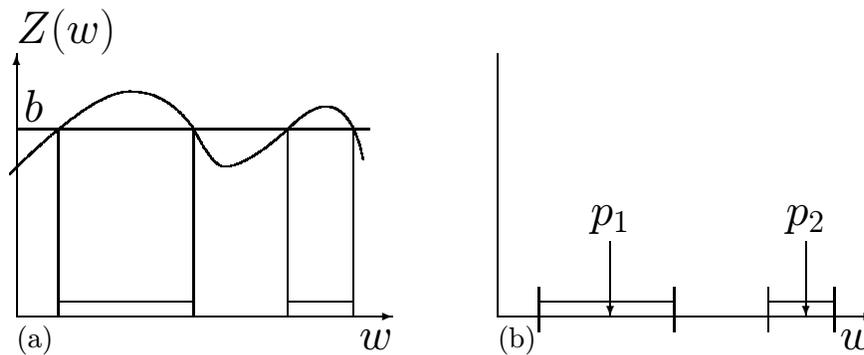
$$\nu_{\mathcal{T}}(w) - \mathcal{E}(w) > \epsilon \quad \leftrightarrow \quad Z(w) > b = \epsilon\sqrt{n}$$

POISSON CLUMPING: GENERALITIES

Viewpoint

Poisson clumping heuristic (PCH) introduced by David Aldous
Tool of wide applicability for estimating exceedance probabilities

$\{w : Z(w) > b\}$ a scattering of ‘clumps’ in \mathcal{W}



Ingredients

- Poisson process of rate $\lambda_b(w)$ generating $\mathcal{P} = \{p\} \subset \mathcal{W}$.
- Clump process across \mathcal{W} choosing clump sets $\mathcal{C}_b(w) \subset \mathcal{W}$.

Mosaic process

$$\mathcal{S}_b := \bigcup_{p \in \mathcal{P}} (p + \mathcal{C}_b(p)) \quad .$$

Assertion: Provided

1. Z has no long-range dependence and
2. the level b is large,

then

$$\mathcal{S}_b \stackrel{\mathcal{D}}{\approx} \{w : Z(w) > b\}$$

PROBABILITIES VIA PCH

Fundamental Equation

Number of clumps $N_b \sim \text{Pois}(\Lambda_b)$ for $\Lambda_b = \int_{\mathcal{W}} \lambda_b(w) dw$

Clump volume $C_b(w) = \text{vol}(\mathcal{C}_b(w))$

$$(1) P(Z(w) > b) = \lambda_b(w) EC_b(w).$$

(2) Since N_b is Poisson,

$$P(N_b > 0) = 1 - e^{-\Lambda_b} \simeq \int_{\mathcal{W}} \lambda_b(w) dw = \int_{\mathcal{W}} \frac{P(Z(w) > b)}{EC_b(w)} dw$$

Finally

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)} dw$$

Summary

Result is sum of exceedance probabilities, diminished by a factor accounting for stochastic dependence.

Have pursued this program

$$\begin{array}{ccccc} \text{Empirical Process} & \xrightarrow{\text{FCLT}} & \text{Gaussian Process} & \xrightarrow{\text{PCH}} & \text{Mosaic Process} \\ \nu_{\mathcal{T}}(w) - \mathcal{E}(w) & & Z(w), R(w, v) & & \lambda_b(w), C_b(w) \end{array}$$

SUMMARY

Qualitative Picture

- Mismatches of size b between $\nu_{\mathcal{T}}(w)$ and $\mathcal{E}(w)$ occur based on a PP of intensity λ
- Overwhelmingly probable that such a discrepancy will occur at a variance maximum

[To Be Seen]

- The size of the region of mismatch varies inversely with b
- In the Gaussian case, the shape of the area depends on $R(w, v)$ for $w \approx v$

Now find clump shape and size:

- Allow probability estimates for assessing generalization
- Qualitative understanding of variability of error surface

SMOOTH ACTIVATIONS: CLUMP SIZE

Activations have two derivatives $\implies Z(w)$ has two derivatives:

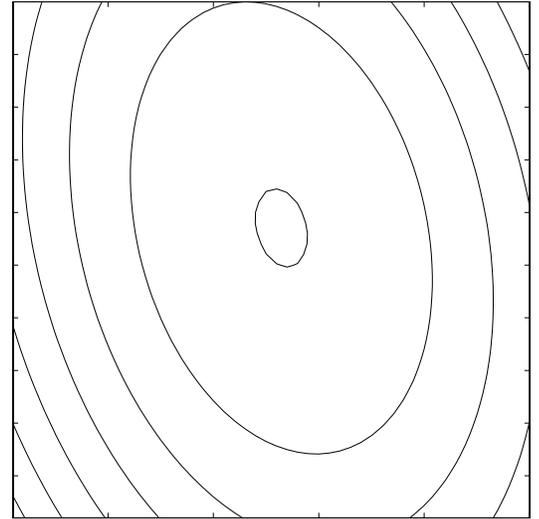
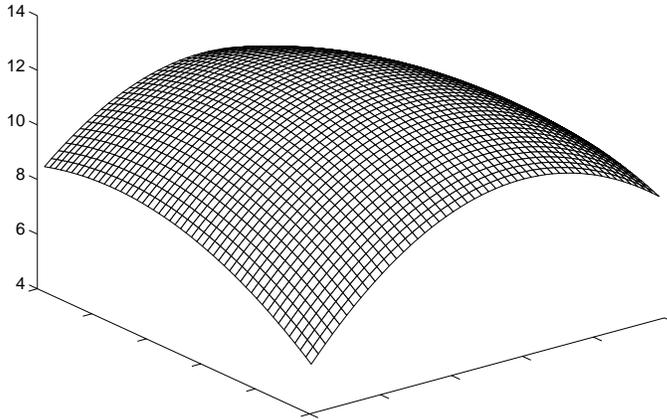
$$Z(w) \simeq Z_0 + (w - w_0)^\top G + \frac{1}{2}(w - w_0)^\top \mathbf{H}(w - w_0)$$

Gradient $G = \nabla Z(w)$ and Hessian $\mathbf{H} = \nabla \nabla Z(w)$

Downward-turning parabola peaks near w_0 and attains height $\geq b$

Clump size is volume V of an ellipsoid in R^d

$$V = \kappa_d \frac{(2(Z_0 - b) - G^\top \mathbf{H}^{-1} G)^{d/2}}{|-\mathbf{H}|^{1/2}} .$$



Since \mathbf{H} and $Z(w_0)$ are jointly Gaussian

$$E[\mathbf{H} \mid Z(w_0) = z] = \frac{-z}{\sigma^2(w_0)} \Lambda_{02}(w_0)$$

$$\begin{aligned} \Lambda_{02}(w_0) &= -E Z(w_0) \mathbf{H} \\ &= -\nabla_w \nabla_w R(w_0, w)|_{w=w_0} \end{aligned}$$

SMOOTH ACTIVATIONS: RESULTS

Clump Size

$$EC_b(w) \simeq E[V \mid Z(w_0) > b] \simeq (2\pi)^{d/2} \frac{(\sigma(w)/b)^d}{|\Lambda_{02}(w)/\sigma^2(w)|^{1/2}}$$

- Correct invariance to scale change
- Local shape determined by $\Lambda_{02}(w)$
- Size diminishes as $1/b$ per dimension

Exceedance Probability

For a unique variance maximum, PCH estimates

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \frac{|\Lambda_{02}(\bar{w})|^{1/2}}{|\Lambda_{02}(\bar{w}) - \Lambda_{11}(\bar{w})|^{1/2}} \bar{\Phi}(b/\bar{\sigma})$$

Also $\Lambda_{02} - \Lambda_{11} = -\nabla\nabla\sigma^2(w)/2 > 0$ at $w = \bar{w}$

Leading constant > 1

Gives $O((v/n)^{1/2})$ shrinking

Normalized Exceedance Probability

$$P\left(\left\|\frac{Z(w)}{\sigma(w)}\right\|_{\mathcal{W}} > b\right) \simeq (2\pi)^{d/2} b^d \bar{\Phi}(b) \int_{\mathcal{W}} \left|\frac{\Lambda_{11}}{\sigma^2(w)}\right|^{1/2} dw$$

where

$$\Lambda_{11}(w_0) = EGG^T = \nabla_w \nabla_v R(w, v)|_{w=v=w_0}$$

Gives $O((v/n)^{1/2})$ shrinking of normalized discrepancy

PERCEPTRON EXAMPLE

- Data $x \in R^p$ has rotationally symmetric distribution
- Networks $\eta(x; w) = 1_{[0, \infty)}(w^\top x)$
WLOG $w^\top w = 1$
 \mathcal{W} = surface of unit ball in R^{d+1} (d free weights)
- Let $y = \eta(x; w^0)$
Problem invariant w.r.t rotations around the axis w^0 .

Clump Size

Easily $R(w, w') = 1/4 - (2\pi)^{-1}|w' - w|_2 + O(|w' - w|_2^2)$

By conditioning on $Z(w) > b$, rescaling w -axis, and exploiting similarity to a canonical process, find

$$EC_b(w) = (1/K_{d,1})(\pi/8)^d / b^{2d}$$

Exceedance Probability

$$\begin{aligned} P(\|Z(w)\|_{\mathcal{W}} > b) &\simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)} dw \\ &\vdots \\ &\simeq \frac{\pi}{4} \frac{8^d}{\pi^{d/2} \Gamma(d/2)} K_{d,1} b^{2d-2} e^{-2b^2} \end{aligned}$$

Symmetry allows reduction to 1-dimensional integral, which is approximated by Laplace's method.

The shrinkage rate is

$$\left(\frac{1.3d}{n}\right)^{1/2} < \epsilon < \left(\frac{(1.36 + (1/3) \log d)d}{n}\right)^{1/2}$$

SUMMARY OF PCH RESULTS

Clump Sizes

- Smooth process: $E C_b(w) = |\Lambda_{02}(\bar{w})|^{-1/2} / b^d$
- Isotropic rough process: $E C_b(w) = K_{d,1}^{-1}(1/b^{2d})$
- Separable rough process: $E C_b(w) = 1/b^{2d}$

Two qualitatively different regimes

Shrinkage Rates

- Learning smooth functions: $\epsilon = (Kd/n)^{1/2}$
- Learning with perceptron in R^d : $\epsilon \simeq (1.3d/n)^{1/2}$
- Learning with orthants in R^d : $\epsilon = (d/n)^{1/2}$
- Learning with rectangles in R^d : $\epsilon = (2d/n)^{1/2}$

Architecture/distribution-sensitive rates

No $\log(n/d)$ factor

EMPIRICAL COUNTERPARTS

Mean bundle size

$$\begin{aligned} E[D_b | Z(w) > b] &= E[\text{vol}\{w' : Z(w') > b\} | Z(w) > b] \\ &= \int_{\mathcal{W}} P(Z(w') > b | Z(w) > b) dw' \end{aligned}$$

Total exceedance volume provided exceedance at w .

A Simple Calculation

$Z(w)$ and $Z(w')$ are jointly normal. Define

$$\begin{aligned} \zeta &= \zeta(w, w') := (\sigma/\sigma') \frac{1 - \rho\sigma'/\sigma}{\sqrt{1 - \rho^2}} \\ &= \left(\frac{1 - \rho}{1 + \rho} \right)^{1/2} \quad (\text{if } \sigma \text{ constant}) \end{aligned}$$

If $b/\sigma \gg 1$ and $\rho\sigma/\sigma' \leq 1$,

$$E[D_b | Z(w) > b] \simeq \int_{\mathcal{W}} \bar{\Phi}((b/\sigma)\zeta) dw'$$

Harmonic Mean Inequality

If $Z(w)$ continuous and $D_b < \infty$ a.s.,

$$\begin{aligned} P(\|Z(w)\|_{\mathcal{W}} > b) &= \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{E[D_b^{-1} | Z(w) > b]^{-1}} dw \\ &\geq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{E[D_b | Z(w) > b]} dw \end{aligned}$$

The bound differs from the asymptotic ($b \rightarrow \infty$) value by constant factors.

CONCLUSIONS

Have examined the discrepancy between the (stochastic) error surface and its (deterministic) mean, the true error

Are interested in large values of this discrepancy — they have implications for generalization

Considering the discrepancy as a stochastic process, have pursued two distributional approximations to it...

1. Normal approximation (provided don't go into tails)
2. The Poisson clumping heuristic

PCH says large discrepancies are scattered independently throughout weight space via a PP.

Shape of the clumps given by correlation structure of generating process

Size depends inversely on level of discrepancy

Can develop close approximations to the true exceedance probability

Easier-to-compute analogs to the clump size can be developed

`mjt@aig.jpl.nasa.gov`

`http://www-aig.jpl.nasa.gov/home/mjt/`