Lecture ~~9~~ 10 Generalizaiton

IEMS 402 Statistical Learning

Northwestern

Rademacher Complexity

Definition. The *empirical Rademacher complexity* of $\mathcal{F}$ is defined to be

$$\hat{R}_m(\mathcal{F}) = \mathsf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right) \right]$$

where $\sigma_1, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$. We will refer to such random variables as *Rademacher variables*.

VC Dimension

Definition 3 (Growth Function). *For any natural number m, define,*

$$\Pi_C(m) = \max\{\underbrace{|\Pi_C(S)|}_{\leq 2^m} \mid |S| = m\}$$

data

$\{\pm 1\}^m$

different ways of hypothesis space to classify the data

Hypothesis space: functions maps from data to ± 1

m data $\rightarrow \{\pm 1\}^m$

Thm. Rademacher Complexity < Growth Function

different numbers of output in $\{\pm 1\}^m$

VC Dimension

Definition 3 (Growth Function). *For any natural number m, define,*

$$\Pi_C(m) = \max\{\boxed{|\Pi_C(S)|} \mid |S| = m\}$$

different ways of hypothesis space to classify the data

Grow at a polynomial at degree d

Can be bounded boun thar largest m such that $\Pi_C(m) = 2^m$

VC Dimension d

Generalization Measures

We investigate more then 40 complexity measures taken from both theoretical bounds and empirical studies. We train over 10,000 convolutional networks by systematically varying commonly used hyperparameters.

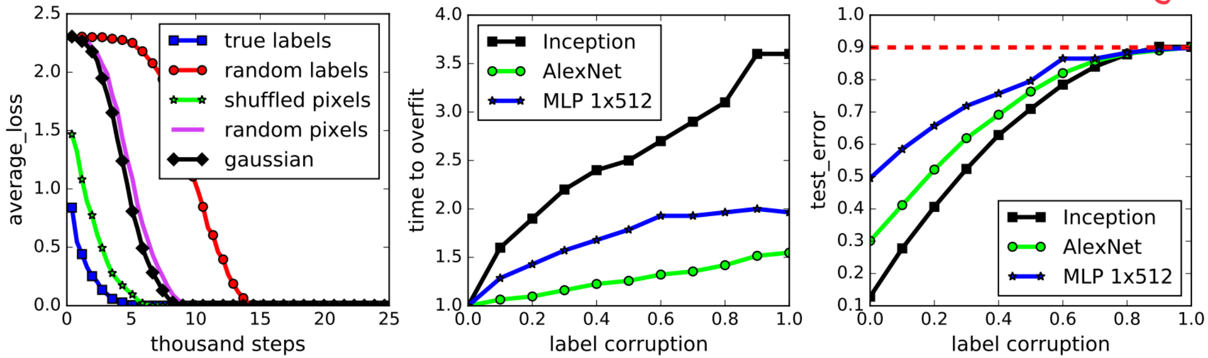
		batch size	dropout	learning rate	depth	optimizer	weight decay	width	overall τ	Ψ
Corr	vc dim 19	0.000	0.000	0.000	-0.909	0.000	0.000	-0.171	-0.251	-0.154
	# params 20	0.000	0.000	0.000	-0.909	0.000	0.000	-0.171	-0.175	-0.154
	$1/\gamma$ (22)	0.312	-0.593	0.234	0.758	0.223	-0.211	0.125	0.124	0.121
	entropy 23	0.346	-0.529	0.251	0.632	0.220	-0.157	0.104	0.148	0.124
	cross-entropy 21	0.440	-0.402	0.140	0.390	0.149	0.232	0.080	0.149	0.147
	oracle 0.02	0.380	0.657	0.536	0.717	0.374	0.388	0.360	0.714	0.487
	oracle 0.05	0.172	0.375	0.305	0.384	0.165	0.184	0.204	0.438	0.256
	canonical ordering	0.652	0.969	0.733	0.909	-0.055	0.735	0.171	N/A	N/A
									$\lvert\mathcal{S}\rvert = 2$	$\min \forall \lvert\mathcal{S}\rvert$
MI	vc dim	0.0422	0.0564	0.0518	0.0039	0.0422	0.0443	0.0627	0.00	0.00
	# param	0.0202	0.0278	0.0259	0.0044	0.0208	0.0216	0.0379	0.00	0.00
	$1/\gamma$	0.0108	0.0078	0.0133	0.0750	0.0105	0.0119	0.0183	0.0051	0.0051
	entropy	0.0120	0.0656	0.0113	0.0086	0.0120	0.0155	0.0125	0.0065	0.0065
	cross-entropy	0.0233	0.0850	0.0118	0.0075	0.0159	0.0119	0.0183	0.0040	0.0040
	oracle 0.02	0.4077	0.3557	0.3929	0.3612	0.4124	0.4057	0.4154	0.1637	0.1637
	oracle 0.05	0.1475	0.1167	0.1369	0.1241	0.1515	0.1469	0.1535	0.0503	0.0503
	random	0.0005	0.0002	0.0005	0.0002	0.0003	0.0006	0.0009	0.0004	0.0001

Table 1: Numerical Results for Baselines and Oracular Complexity Measures

Jiang, Yiding, et al. "Fantastic generalization measures and where to find them." arXiv preprint arXiv:1912.02178

Why?

y=x means perfect generalization

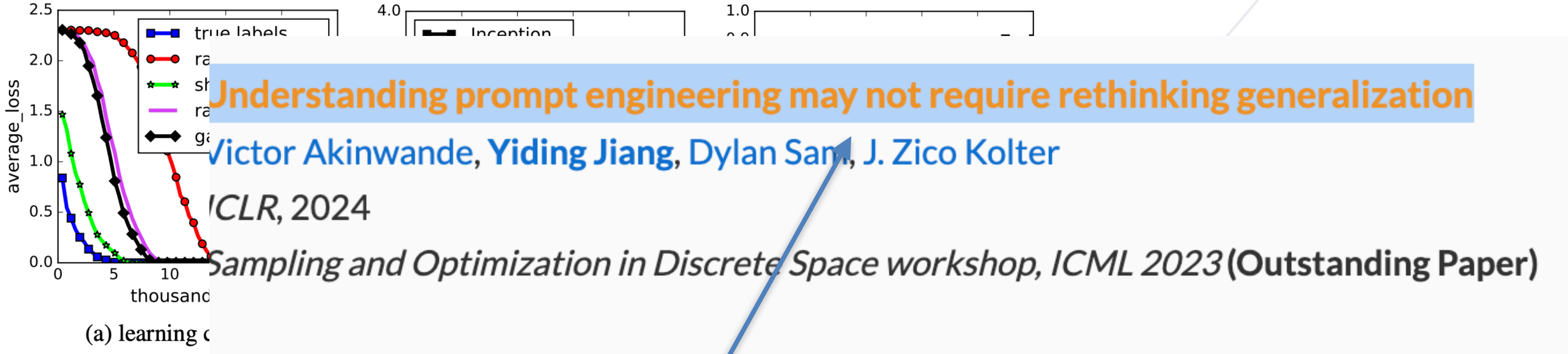


(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

Zhang, Chiyuan, et al. "Understanding deep learning (still) requires rethinking generalization." *Communications of the ACM* 64.3 (2021): 107-115.

Why?



Understanding prompt engineering may not require rethinking generalization

Victor Akinwande, Yiding Jiang, Dylan Sam, J. Zico Kolter

ICLR, 2024

Sampling and Optimization in Discrete Space workshop, ICML 2023 **(Outstanding Paper)**

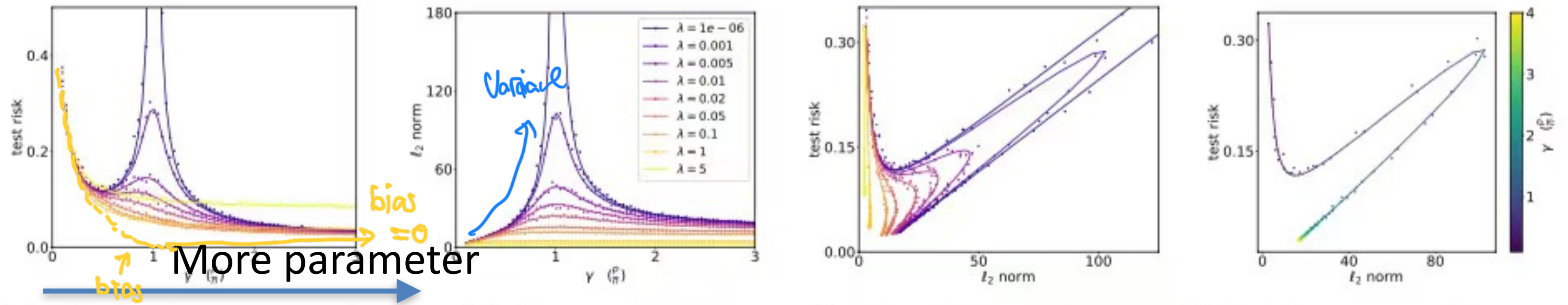
Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

Zhang, Chiyuan, et al. "Understanding deep learning (still) requires rethinking generalization." *Communications of the ACM* 64.3 (2021): 107-115.

Norm Matters

Linear regression

https://arxiv.org/abs/2502.01585

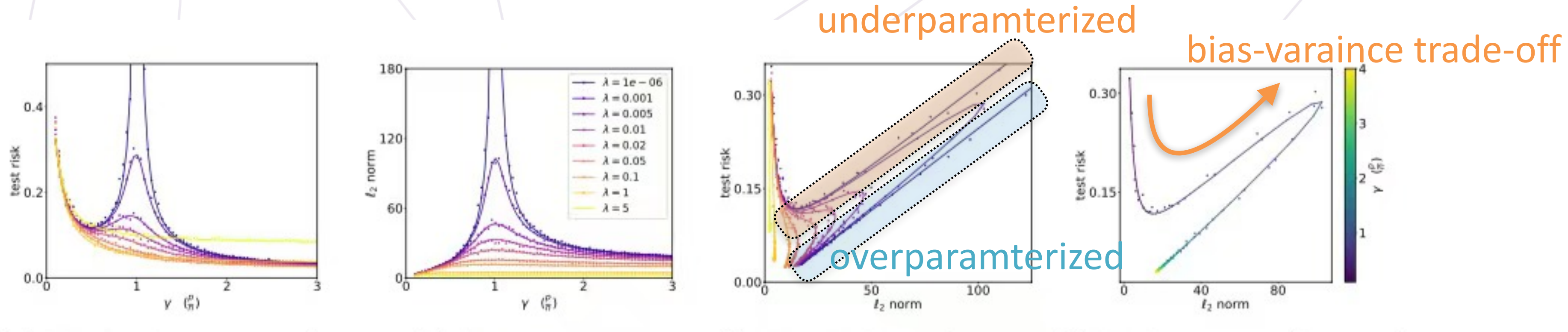


(a) Test Risk vs. $\gamma := p/n$ (b) ℓ_2 norm vs. γ (c) Test Risk vs. ℓ_2 norm (d) Risk vs. norm (λ=0.001)

Neural networks

Bartlett, P.L., 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE transactions on Information Theory, 44(2), pp.525-536

Norm Matters



(a) Test Risk vs. $\gamma := p/n$ (b) ℓ_2 norm vs. γ (c) Test Risk vs. ℓ_2 norm (d) Risk vs. norm (λ=0.001)

https://arxiv.org/abs/2502.01585

Re-examining Double Descent and Scaling Laws under Norm-based Capacity via Deterministic Equivalence

$\forall \theta \in \Theta$

$\exists \theta_i$ such that $\|\theta_i - \theta\| \leq \varepsilon$

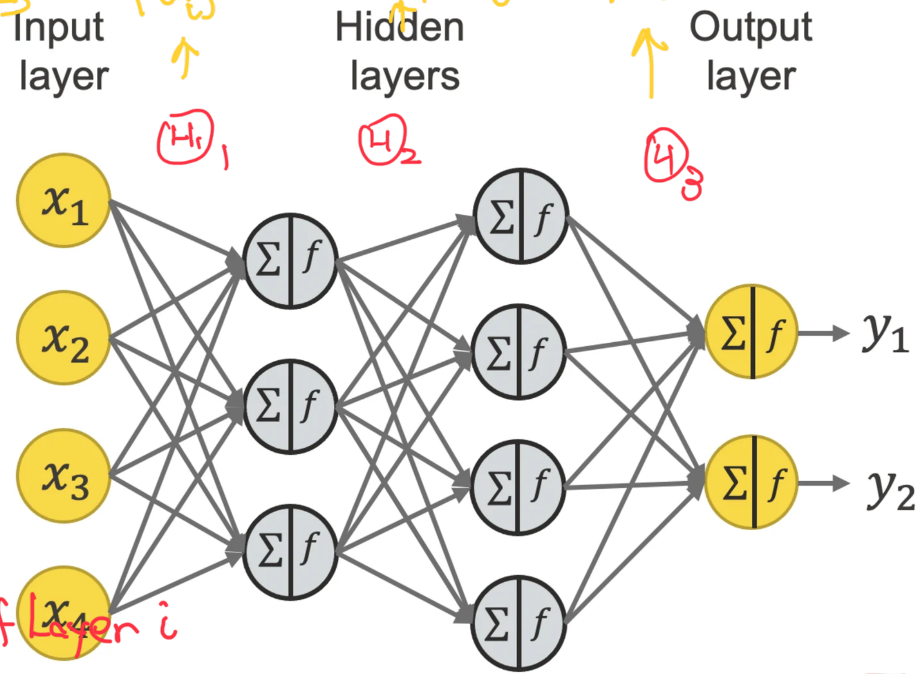
$\Rightarrow \{\theta_i\}$ is a ε-cover.

Covering Number

log-Covering number of NN

ε- Covers → $\{\theta_i^1\}$ $\{\theta_i^2\}$ $\{\theta_i^3\}$

H_1 H_2 H_3



Claim:

$\{(\theta_i^1, \theta_j^2, \theta_k^3), i, j, k\}$

⇒

Covering number of NN

$= \prod_{i=1}^{n}$ Covering number of Layer i

⇒ log Covering number $= \sum_{i=1}^{n}$ log Covering number of layer i $= \sum \|w_i\|$

Rademacher Complexity

Example: Linear

Theorem 4. *A linear hypothesis class* $\mathcal{H}$ *such that* $\forall h \in \mathcal{H}$, $h_w(x) =< w, x > \cancel{\in [-1, 1]}$, *where* $w \in \mathbb{R}^n$ $\|w\|_2 \leq \mathcal{B}$, *and* $x \in \mathbb{R}^n$, $\|x\|_2 \leq \mathcal{X}$, *we have*

$$\hat{\mathcal{R}}_m(\mathcal{H}, \mathcal{S}) \leq \frac{2\mathcal{B}\mathcal{X}}{\sqrt{m}} \tag{9}$$

https://courses.cs.washington.edu/courses/cse522/11wi/scribes/lecture6.pdf

$$
\begin{aligned}
&= \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{\|w\|_2 \le \mathcal{B}} \sum_{i=1}^{m} \sigma_i < w, x_i > \\
&= \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{\|w\|_2 \le \mathcal{B}} < w, \sum_{i=1}^{m} \sigma_i x_i > \\
&\le \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{\|w\|_2 \le \mathcal{B}} \|w\| \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \quad \text{(CauchySchwarz inequality)} \\
&= \frac{2\mathcal{B}}{m}\mathsf{E}_{\vec{\sigma}} \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \\
&= \frac{2\mathcal{B}}{m}\mathsf{E}_{\vec{\sigma}} \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i \sigma_j < x_i, x_j >} \quad \text{(linearity of inner product)} \\
&\le \frac{2\mathcal{B}}{m} \sqrt{\mathsf{E}\sum_{ij} \sigma_i \sigma_j < x_i, x_j >} \quad \text{(Jensen's inequality)} \\
&= \frac{2\mathcal{B}}{m} \sqrt{\sum_{ij} < x_i, x_j > \mathsf{E}\sigma_i \sigma_j} \\
&\le \frac{2\mathcal{B}}{m} \sqrt{\sum_{i} \|x_i\|^2} \\
&\le \frac{2\mathcal{B}}{m} \sqrt{m}\mathcal{X}
\end{aligned}
$$

Fact

Talagrand Contraction

Theorem 2. *If the loss function is λ–Lipschitz, we have*

$$\mathcal{R}_m(l \circ \mathcal{H}) \leq \lambda \mathcal{R}_m(\mathcal{H}) \quad (4)$$

$$(5)$$

$$l \circ H = \{ l(h(x)) : h \in H \}$$

$$\prod_{i=1}^{L} \|W_i\|$$

Different norms…

		batch size	dropout	learning rate	depth	optimizer	weight decay	width	overall τ	Ψ
Corr	Frob distance 40	-0.317	-0.833	-0.718	0.526	-0.214	-0.669	-0.166	-0.263	-0.341
	Spectral orig 26	-0.262	-0.762	-0.665	-0.908	-0.131	-0.073	-0.240	-0.537	-0.434
	Parameter norm 42	0.236	-0.516	0.174	0.330	**0.187**	0.124	-0.170	0.073	0.052
	Path norm 44	0.252	**0.270**	0.049	**0.934**	0.153	0.338	**0.178**	**0.373**	**0.311**
	Fisher-Rao 45	**0.396**	0.147	**0.240**	-0.553	0.120	**0.551**	0.177	0.078	0.154
	oracle 0.02	0.380	0.657	0.536	0.717	0.374	0.388	0.360	0.714	0.487
									$\lvert\mathcal{S}\rvert = 2$	$\min \forall \lvert\mathcal{S}\rvert$
MI	Frob distance	0.0462	0.0530	0.0196	**0.1559**	0.0502	0.0379	0.0506	0.0128	0.0128
	Spectral orig	**0.2197**	**0.2815**	**0.2045**	0.0808	**0.2180**	**0.2285**	**0.2181**	**0.0359**	**0.0359**
	Parameter norm	0.0039	0.0197	0.0066	0.0115	0.0064	0.0049	0.0167	0.0047	0.0038
	Path norm	0.1027	0.1230	0.1308	0.0315	0.1056	0.1028	0.1160	0.0240	0.0240
	Fisher Rao	0.0060	0.0072	0.0020	0.0713	0.0057	0.0014	0.0071	0.0018	0.0013
	oracle 0.05	0.1475	0.1167	0.1369	0.1241	0.1515	0.1469	0.1535	0.0503	0.0503

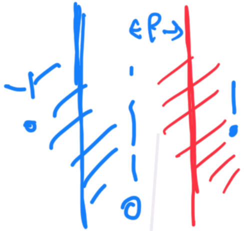
Table 2: Numerical Results for Selected (Norm & Margin)-Based Complexity Measures

Margin

Margin Bounds

Theorem 1. *Let $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ and fix $\rho > 0$, $\delta > 0$. With probability at least $1-\delta$, for all $f \in \mathcal{F}$*

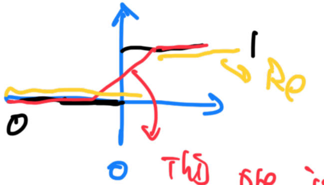
Classify the data to $-1, 1$



$yf(x_i) \geq 0$

$$R(f) \leq \widehat{R}_{\rho}(f) + \frac{2}{\rho}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

0-1 loss is not Lipschitz



training points misclassified or correctly classified with a "confidence"

$$\widehat{R}_{L_\rho}(f) = \frac{1}{n}\sum_{i=1}^{n}\phi_\rho(y_i f(x_i)) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{y_i f(x_i) \leq \rho\} =: \widehat{R}_\rho(f).$$

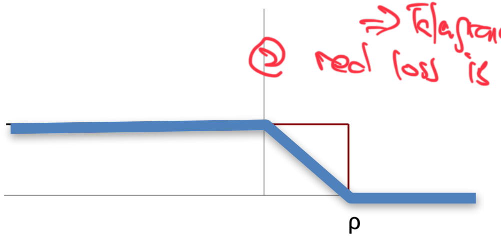
① red loss is Lip

⇒ Talagrand Contraction

This one is Lipschitz

Lip constant is $\frac{2}{\rho}$

② real loss is always larger then R_ρ. ⇒ hinge loss.



Key Insight: "smoothed loss is Lipschitz"

For Neural Network

Wei, Colin, and Tengyu Ma. "Improved sample complexities for deep networks and robust classification via an all-layer margin." arXiv preprint arXiv:1910.04284 (2019).

Algorithm Stability

Stability

notation: S training set, $S^{i,z}$ training set obtained replacing the i-th example in S with a new point $z = (x, y)$. $\beta = O\left(\frac{1}{n}\right)$

Definition

We say that an algorithm $\mathcal{A}$ has **uniform stability** β (is β-stable) if

$$\forall (S, z) \in \mathcal{Z}^{n+1},\ \forall i,\ \sup_{z' \in Z} |V(f_S, z') - V(f_{S^{i,z}}, z')| \leq \beta.$$

https://www.mit.edu/~9.520/spring09/Classes/class09_stability.pdf

What's the result like

Most of the cases $\beta = \frac{k}{n}$, then the generalization bound gives by

with probability $1 - \delta$,

$$I[f_S] \leq I_S[f_S] + \frac{k}{n} + (2k + M)\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Proof Idea: McDiarmid's Inequality

Train Faster, Generalize Better?

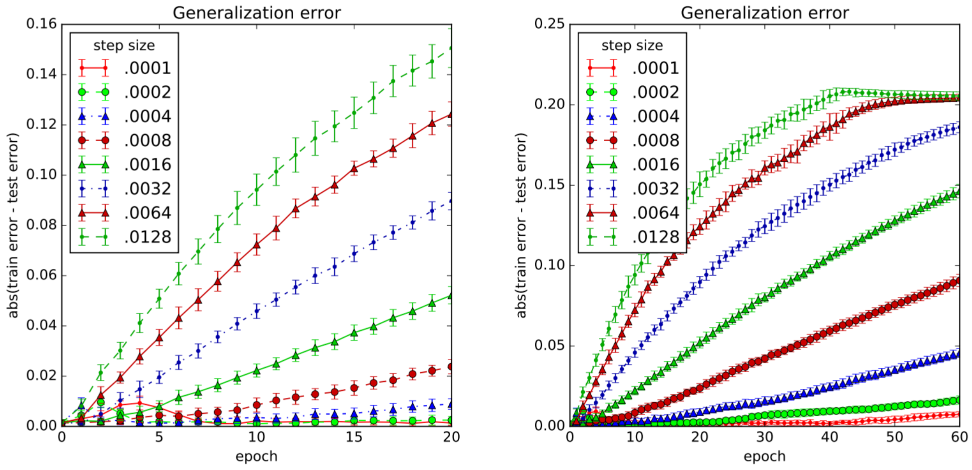


Figure 1: Generalization error as a function of the number of epochs for varying step sizes on Cifar10. Here generalization error is measured with respect to *classification accuracy*. Left: 20 epochs. Right: 60 epochs.

A Trade-off?

Chen, Yuansi, Chi Jin, and Bin Yu. "Stability and convergence trade-off of iterative optimization algorithms." *arXiv preprint arXiv:1804.01619* (2018).

Hardt, Moritz, Ben Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent." International conference on machine learning. PMLR, 2016.

PAC Bayes

Randomized Classifier

We will consider the binary classification task with an input space $\mathcal{X}$ and label set $\mathcal{Y} = \{+1, -1\}$. Let $\mathcal{D}$ be the (unknown) true on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ be a hypothesis class of functions $f : \mathcal{X} \mapsto \mathcal{Y}$. Let $\mathcal{P}$ be the space of probability distributions on $\mathcal{H}$. We consider $0, 1$-valued loss functions $l : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \mapsto \{0, 1\}$.

Definition 1. *Let $Q \in \mathcal{P}$. Define:*

$$\textit{Risk of } Q \; l(Q; \mathcal{D}) = E_{(x,y)\sim\mathcal{D}} \boxed{E_{h\sim Q}} [l(h; (x, y))]$$

$$\textit{Emperical Risk of } Q \; l(Q; D) = \frac{1}{|D|} \sum_{(x,y)\in D} \boxed{E_{h\sim Q}} [l(h; (x, y))]$$

Q not necessary to be the posterior!

PAC-Bayes Bound

Theorem 2. *(McAllester) $\forall \mathcal{D}, \forall \mathcal{H} \forall P \in \mathcal{P} \forall \delta > 0$, we have with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$: $\forall Q \in \mathcal{P}$ (posterior distribution on $\mathcal{H}$ that depends on S),*

Trained Belief

$$\mathrm{KL}\left(l(Q;S) \,\|\, l(Q:\mathcal{D})\right) \leq \frac{\mathrm{KL}\left(Q \,\|\, P\right) + \log\left(\frac{m+1}{\delta}\right)}{m}$$

P should not relates to the data you sampled (S) / Algorithm

but P may be depend on D.

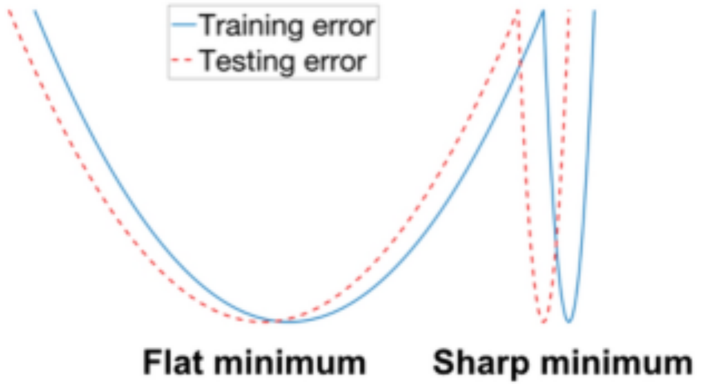
"soft" version of algorithm stability

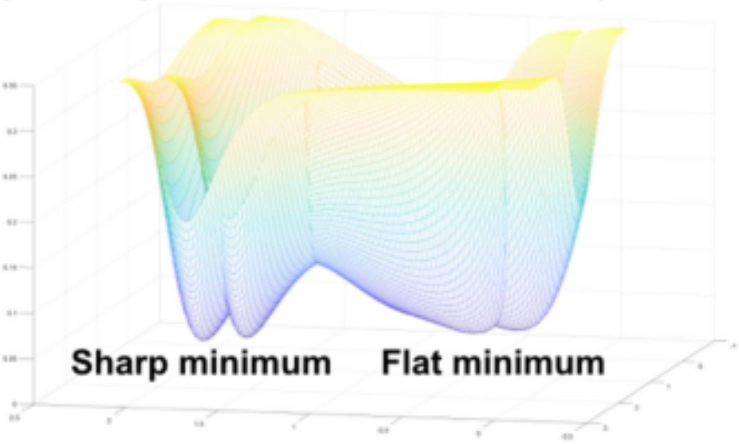
randomized hypothesis.

Basic idea: https://arxiv.org/pdf/2110.11216

$$\log \mathbb{E}_{\theta\sim\pi}\left[e^{h(\theta)}\right] = \sup_{\rho\in\mathcal{P}(\Theta)}\left[\mathbb{E}_{\theta\sim\rho}[h(\theta)] - KL(\rho\|\pi)\right].$$

Sharp Minima





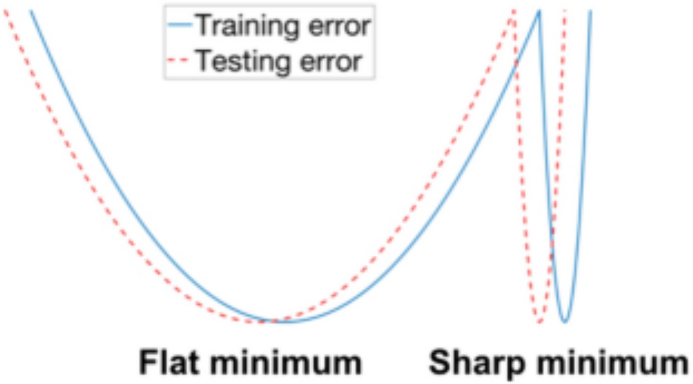
Train with Larger noise

Train with smaller noise

Keskar, Nitish Shirish, et al. "On large-batch training for deep learning: Generalization gap and sharp minima." arXiv preprint arXiv:1609.04836 (2016).

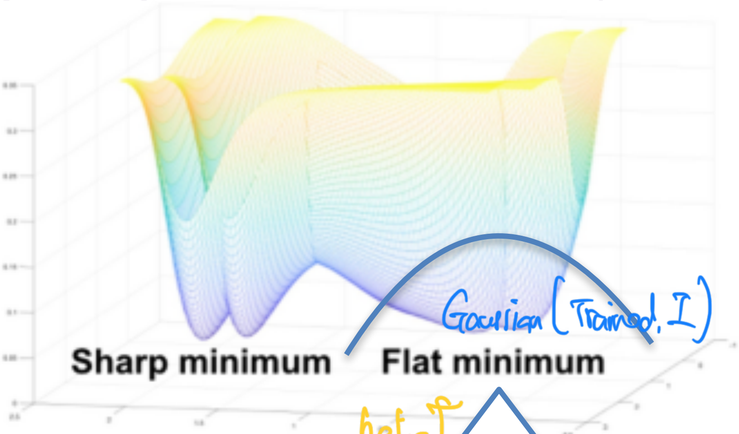
PAC-Bayes?

Neyshabur, Behnam, et al. "Exploring generalization in deep learning."Neurips 2017



Train with Larger noise

Train with smaller noise



before?

PAC Bayes KL(Gaussian(Trained, I) || N(0, I))

Put a distribution over hypothesis

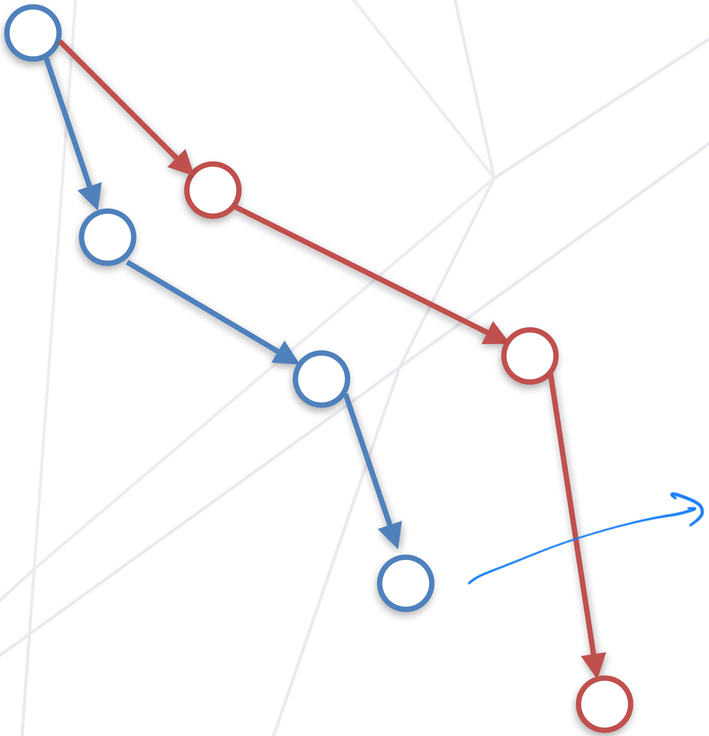
= ||weight||^2

Generalization =

generalization of randomized hypothesis + distance of random and deterministic

norm of weight! sharpness of the minimum!

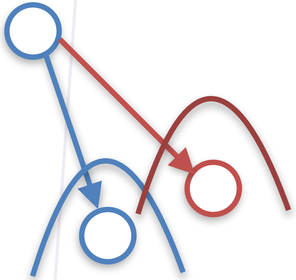
Second Idea



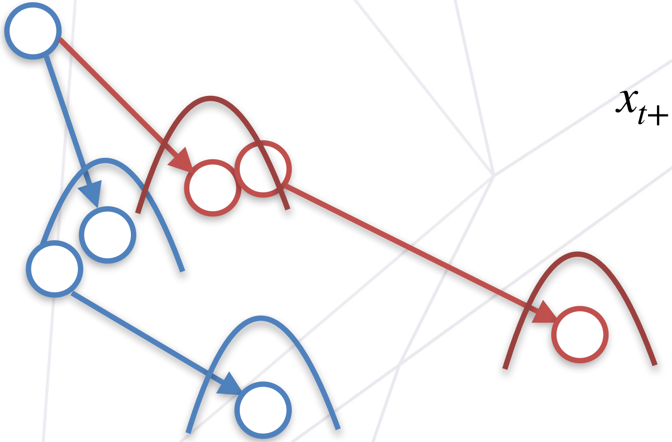
KL divergence is infinite

reason: different support.

Second Idea



Second Idea



Batch size vs. learning rate

Variance to be η

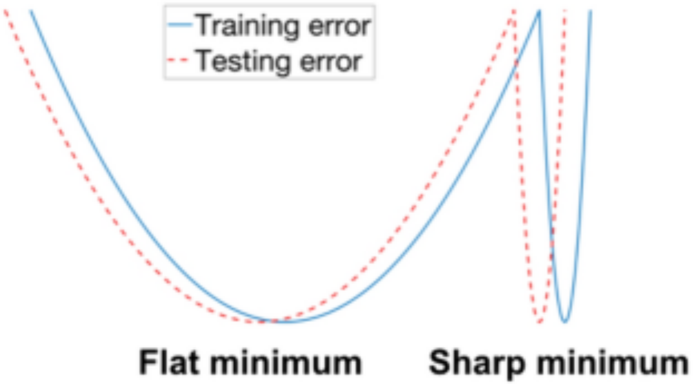
$$x_{t+1} = \underbrace{x_t - \eta \nabla f(x_t)}_{\text{G.D.}} + \underbrace{\sqrt{\eta}\epsilon}_{\text{noise}}, \epsilon \sim N(0,I)$$

$Var(X_1) + Var(X_2) = Var(X_1 + X_2)$

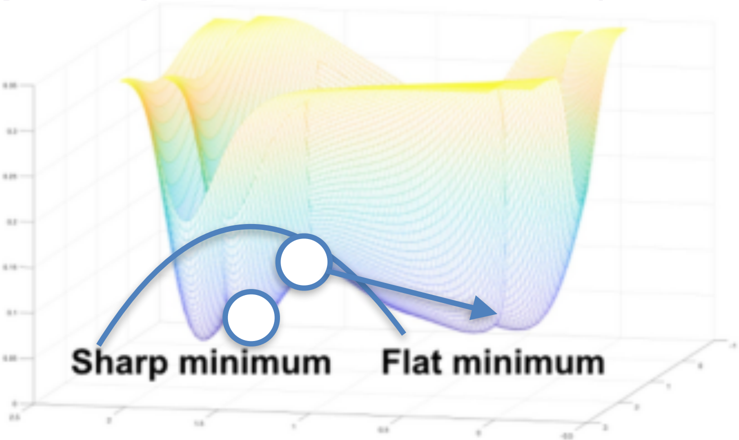
Why $\sqrt{\eta}$?

Mou, Wenlong, et al. "Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints." Conference on Learning Theory. PMLR, 2018.

Sharp Minima



Train with Larger noise Train with smaller noise



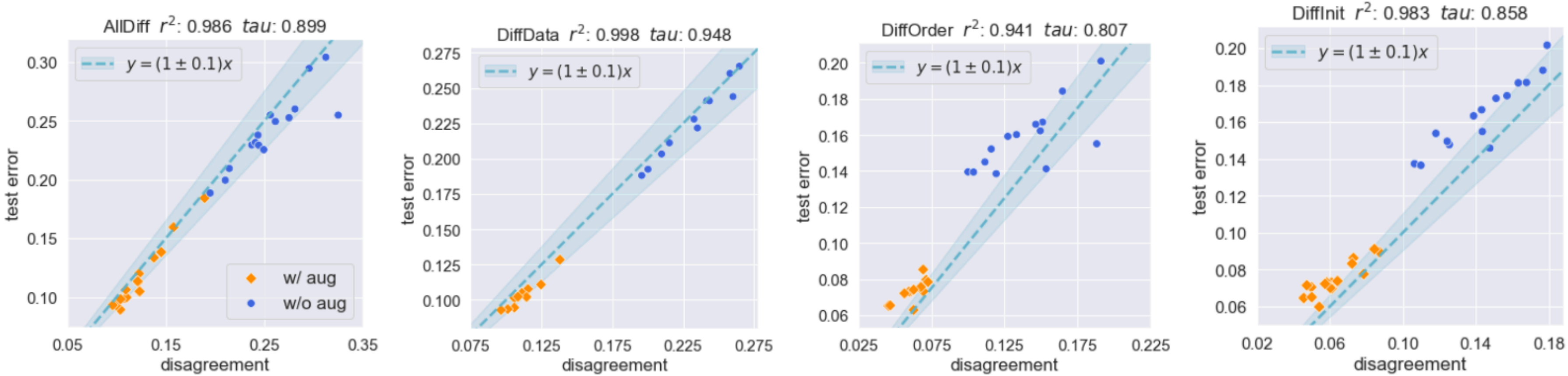
SDE View ⬇️

Smith, Samuel L., and Quoc V. Le. "A bayesian perspective on generalization and stochastic gradient descent." *arXiv preprint arXiv:1710.06451* (2017).

Disaggrement?

Definition 4.1. The stochastic learning algorithm $\mathcal{A}$ satisfies **the Generalization Disagreement Equality (GDE)** on the distribution $\mathscr{D}$ if,

$$\mathbb{E}_{h,h' \sim \mathscr{H}_{\mathcal{A}}}[\mathtt{Dis}_{\mathscr{D}}(h,h')] = \mathbb{E}_{h \sim \mathscr{H}_{\mathcal{A}}}[\mathtt{TestErr}_{\mathscr{D}}(h)]. \quad (3)$$

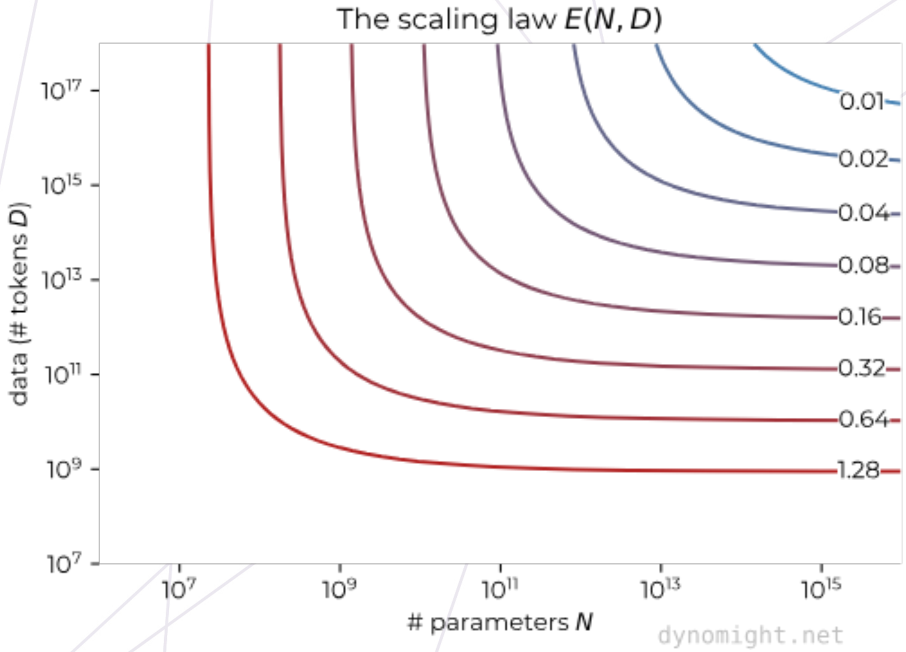


Jiang, Yiding, et al. "Assessing generalization of SGD via disagreement." arXiv preprint arXiv:2106.13799 (2021).

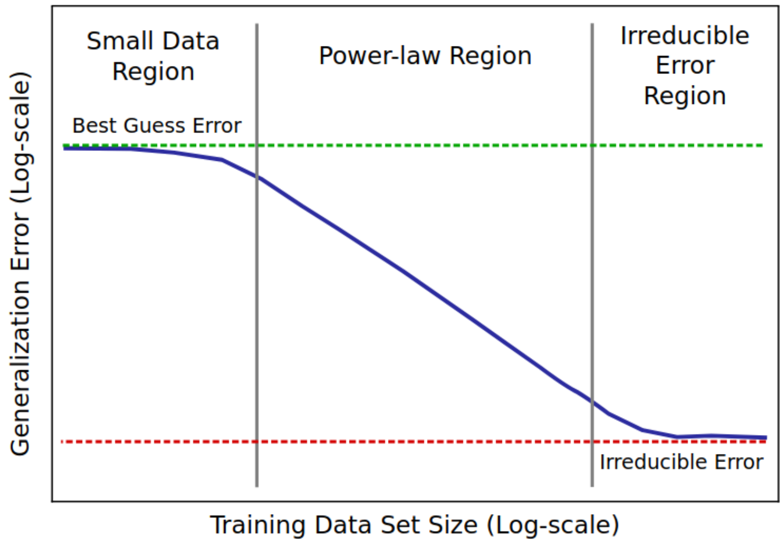
Related works

Angelopoulos A N, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification[J]. arXiv preprint arXiv:2107.07511, 2021.

Scaling Law



The scaling law E(N, D)
data (# tokens D)
parameters N
10^17
10^15
10^13
10^11
10^9
10^7
10^7
10^9
10^11
10^13
10^15
0.01
0.02
0.04
0.08
0.16
0.32
0.64
1.28
dynomight.net



Small Data Region
Power-law Region
Irreducible Error Region
Best Guess Error
Irreducible Error
Generalization Error (Log-scale)
Training Data Set Size (Log-scale)