To Understand Deep Learning We Need to Understand Kernel Learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal

Yihan Zhou(Joey)

Nov 20th, 2019

The University of British Columbia

Mysterious Generalization Behaviour of Deep Models

Deep models, usually heavily over-parametrized, tend to easily fit the training data with 0 training error.

But why can these overfitted models generalized well?

Mysterious Generalization Behaviour of Deep Models

Deep models, usually heavily over-parametrized, tend to easily fit the training data with 0 training error.

But why can these overfitted models generalized well?

This is not unique for deep models, kernel machines have similar generalization performance.

Recap of Classical Bounds

A classical learning paradigm is Empirical Risk Minimization(ERM). Given data $\{(x_i, y_i), i = 1, \cdots, n\}$ sampled from a probability distribution P on $\Omega \times \{1, -1\}$, a class of functions $\mathcal{H} : \Omega \to \mathbb{R}$ and a loss function l, ERM finds a minimizer of the empirical loss:

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} L_{emp}(f) := \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_i l(f(x_i), y_i).$$

Many classical generalization bound are of the form $|\mathbb{E}[l(f^*(x), y)] - L_{emp}(f)| < O^*(\sqrt{c/n})$. The c is a measure of complexity of $\mathcal{H}$, e.g., VC-dimension, Rademacher complexity, covering number, fat shattering dimensions.

Recap of Kernel Learning

Let $K(x, z) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definitely kernel, then there exists a corresponding Reproducing Kernel Hilbet Space(RKHS) $\mathcal{H}$ of functions on $\mathbb{R}^d$, associated to the kernel $K(x, z)$.
The minimum norm interpolent is defined as

$$f^* = \underset{f \in \mathcal{H}, f(x_i) = y_i}{\operatorname{argmin}} \|f\|_{\mathcal{H}}.$$

By Representer Theorem, f^* can be written explicitly as

$$f^*(\cdot) = \sum \alpha_i^* K(x_i, \cdot).$$

Then we can formulate kernel learning as

$$\alpha^* = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{n} l\left(\left(\sum_{j=1}^{n} \alpha_i K(x_j, x_i)\right), y_i\right).$$

Recap of Kernel Learning

This formulation is an unconstrained optimization problem on a finite-dimensional space $\mathbb{R}^n$, so it can be solved by iterative methods.

RKHS norm of a function of the form $f(\cdot) = \sum \alpha_i K(x_i, \cdot)$ can be computed as

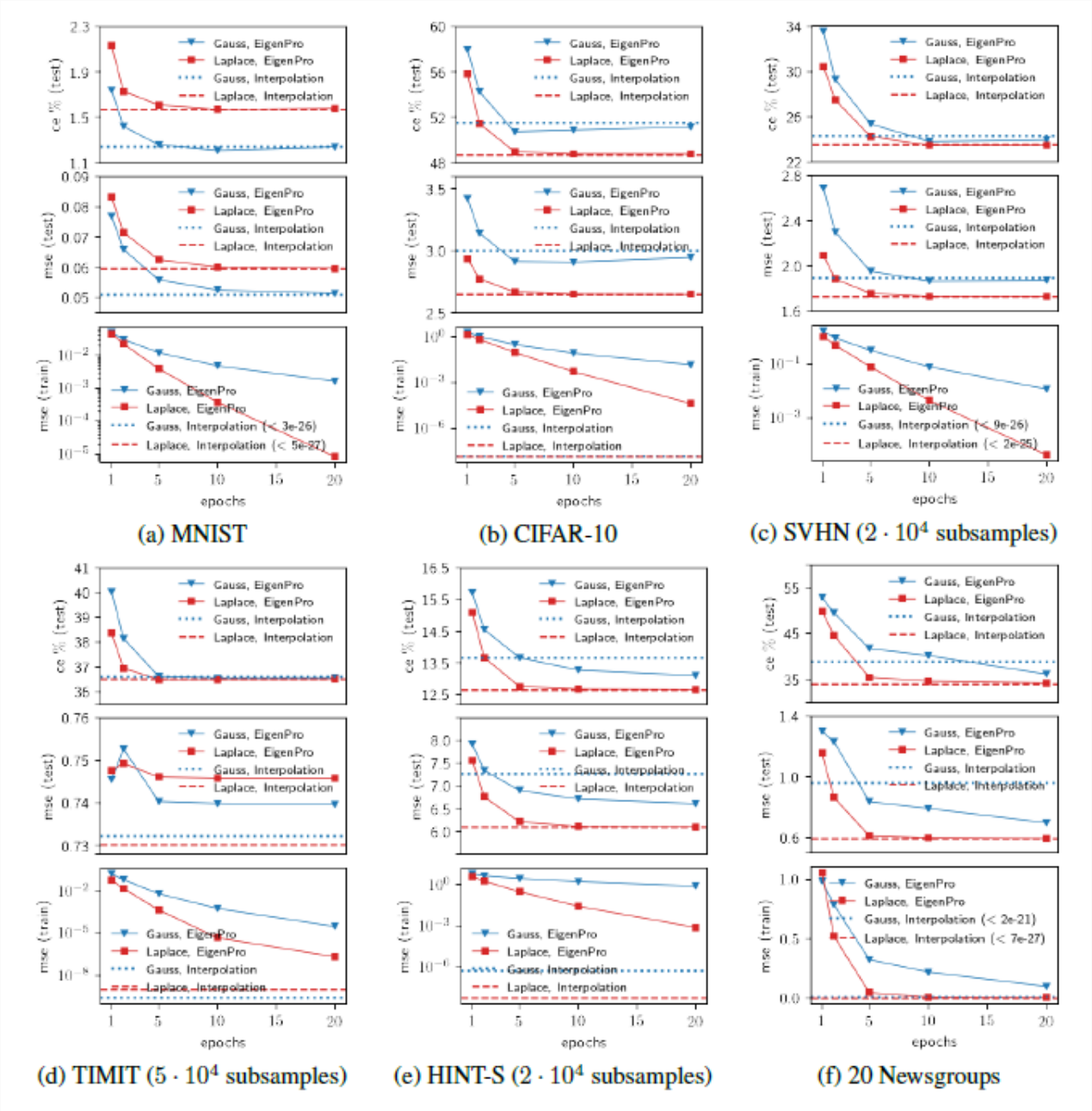
$$\|f\|_{\mathcal{H}}^2 = \langle \alpha, K\alpha \rangle = \sum_{ij} \alpha_i K_{ij} \alpha_j.$$

- Gaussian Kernel: $K(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$
- Laplacian Kernel: $K(x, z) = \exp\left(-\frac{\|x-z\|}{\sigma}\right)$

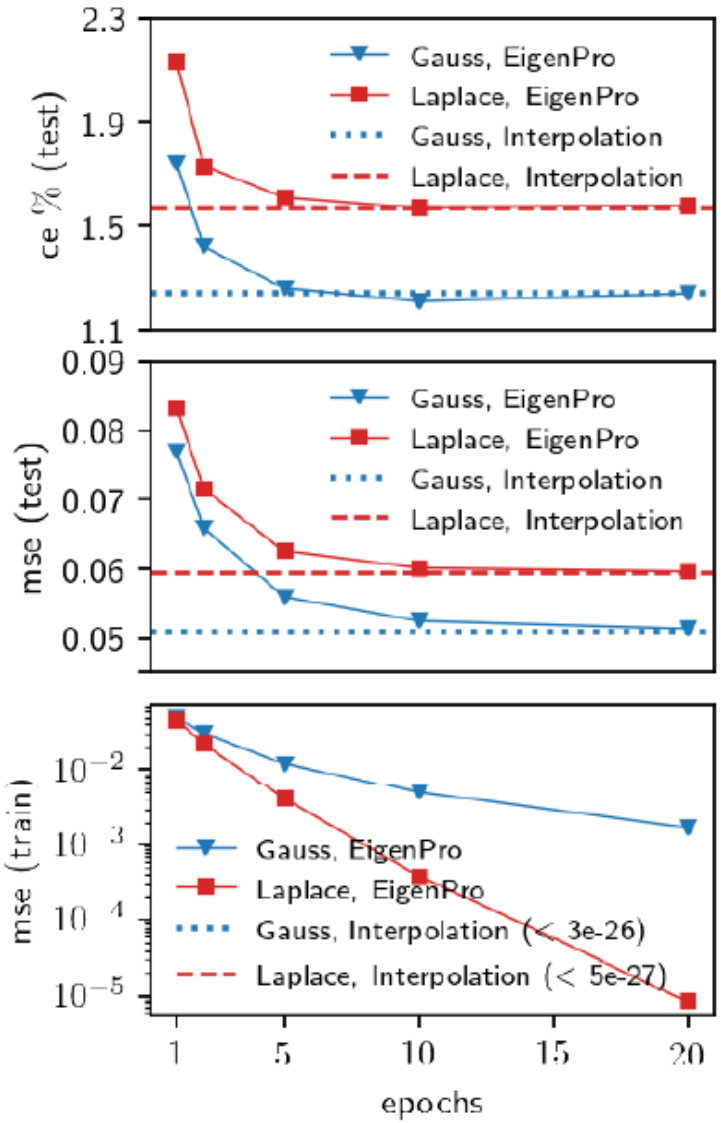
Experiments Setup

- Kernels: Gaussian and Laplacian
- Optimizer:
  - EigenPro-SGD
  - Directly method
- Six dataset: MNIST, CIFAR-10, SVHN, TIMIT, HINT-S, 20 Newsgroups

Experiment Results



Experiment Results



(a) MNIST

- Training square loss(mse)/training classification error approaches zero.
- Test error, both **mse** and **ce**, remains stable, in most cases, keeps decreasing and then stablize.
- Direct solutions always provide a highly accurate interpolation for the training data.
- Interpolated solution on test is either optimal or close to optimal in **mse** and **ce**.

Conclusion

- Generalization performance of overfitted/interpolated kernel classifiers closely parallels behaviors of deep networks.

- It has been observed before that very small values of regularization parameters frequently lead to optimal performances[SSSSC11, TBRS13]. Similar observations were also made for Adaboost and Random Forests[SFB$^+$98].

Existing Generalization Bounds for Kernel Methods

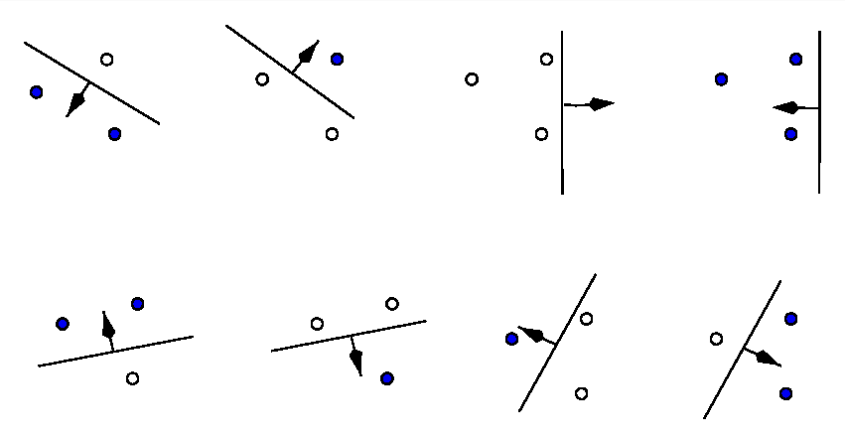
Most of the available bounds for kernel methods (see, e.g., [SC08, RCR15]) are of the following (general) form:

$$\left| \frac{1}{n} \sum_i l(f(x_i), y_i) - \mathbb{E}_P \left[l(f(x), y) \right] \right| \leq C_1 + C_2 \frac{\|f\|_{\mathcal{H}}^{\alpha}}{n^{\beta}}, \quad C_1, C_2, \alpha, \beta \geq 0.$$

Some bounds are potentially logarithmic[Bel18, GK17], but all of them include a non-zero accuracy parameter($\frac{1}{\epsilon}$), so it will not apply to interpolated classifier.

Digression: Fat Shattering Dimension

We say a function class H shatters a set $\mathcal{X} = \{x_1, \cdots, x_n\}$ if it contains all of the possible label assignments of the set.

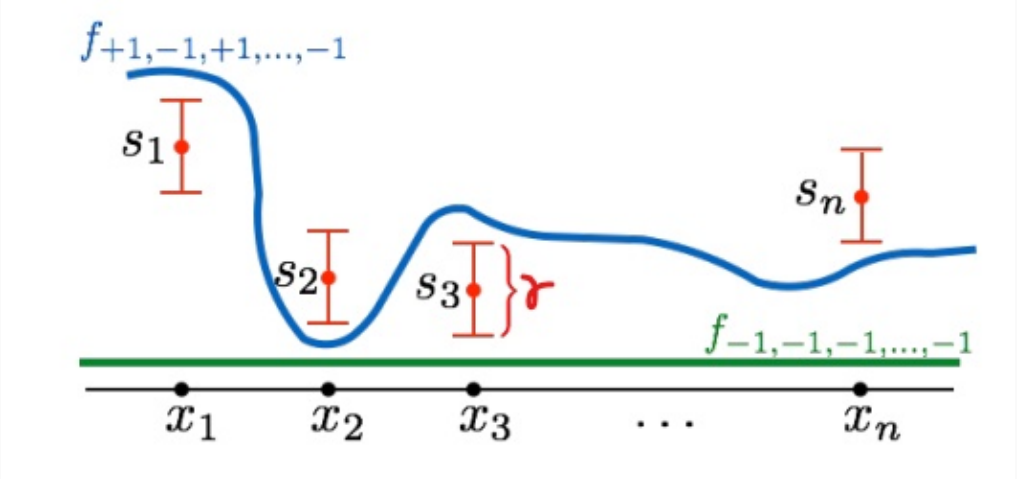


VC dimension is the maximum cardinality of all of the sets can be shattered by H.

Digression: Fat Shattering Dimension

We say that H shatters $x_1 \cdots, x_n$ at scale γ, if there exists witness $s_1, \cdots, s_n$ such that, for every $\epsilon \in \{\pm 1\}^n$, there exists $f \in H$ such that

$$\forall t \in [n], \quad \epsilon_t \cdot (f_\epsilon(x_t) - s_t) \geq \frac{\gamma}{2}.$$



$\text{fat}_\gamma(\mathcal{H}) = \max\{n : \exists x_1, \cdots, x_n \in \mathcal{X} \text{ s.t. } H \text{ } \gamma\text{-shatters } x_1, \cdots, x_n\}$.

Lower Bound for Interpolated Kernel

Assumptions:

- $(x_i, y_i) \in \Omega \times \{-1, 1\}$ be a labelled dataset
- Domain $\Omega \subseteq \mathbb{R}^n$ is bounded
- Bayes optimal classifier(the label noise) is not 0, i.e., y is not a deterministic function of x

Definition: A function $h \in \mathcal{H}$ t-overfits the data, if it achieves zero classification loss, and, additionally, $\forall i \; y_i h(x_i) > t > 0$ for at least a fixed portion of the training data.

This condition is necessary as zero classification loss classifiers with arbitrarily small norm can be obtained by simply scaling any interpolating solution.

Lower Bound for Interpolated Kernel

Theorem: Let $(x_i, y_i), i = 1, \cdots, n$ be data sampled from P on $\Omega \times \{-1, 1\}$. Assume that y is not a deterministic function of x on a subset of non-zero measure. Then, with high probability, any h that t-overfits the data, satisfies

$$\|h\|_{\mathcal{H}} > Ae^{Bn^{1/d}}$$

for some constants $A, B > 0$ depending on t.

This is an exponential lower bound for the Hilbert space norm of any overfitted classifier and it makes the kernel generalization bound trivial.

Proof Sketch

Let $B_R = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} < R\} \subseteq H$ be a ball of radius R in the RKHS $\mathcal{H}$. Let l be the hinge loss with margin t : $l(f(x), y) = (t - yf(x))_+$. By the classical results on fat shattering dimension[AB09], $\exists C_1, C_2 > 0$ such that with high probability $\forall f \in B_R$:

$$\left| \frac{1}{n} \sum_i l(f(x_i), y_i) - \mathbb{E}_P \left[l(f(x), y) \right] \right| \leq C_1 \gamma + C_2 \sqrt{\frac{\text{fat}_\gamma(B_R)}{n}}.$$

Let $h \in B_R$ t-overfits the data, $\frac{1}{n} \sum_i l(f(x_i), y_i) = 0$ and

$$0 < \mathbb{E}_P \left[l(f(x), y) \right] - C_1 \gamma < C_2 \sqrt{\frac{\text{fat}_\gamma(B_R)}{n}}.$$

Proof Sketch

The previous inequality gives

$$\text{fat}_\gamma(B_R) > \frac{n}{C_2}(\mathbb{E}_P\left[l(f(x), y)\right] - C_1\gamma)^2.$$

On the other hand, [Bel18] gives a bound on the fat_γ dimension of the form

$$\text{fat}_\gamma(B_R) < O\left(\log^d\left(\frac{R}{\gamma}\right)\right).$$

Combining these two together, we get the desired bound

$$R > Ae^{Bn^{1/d}}.$$

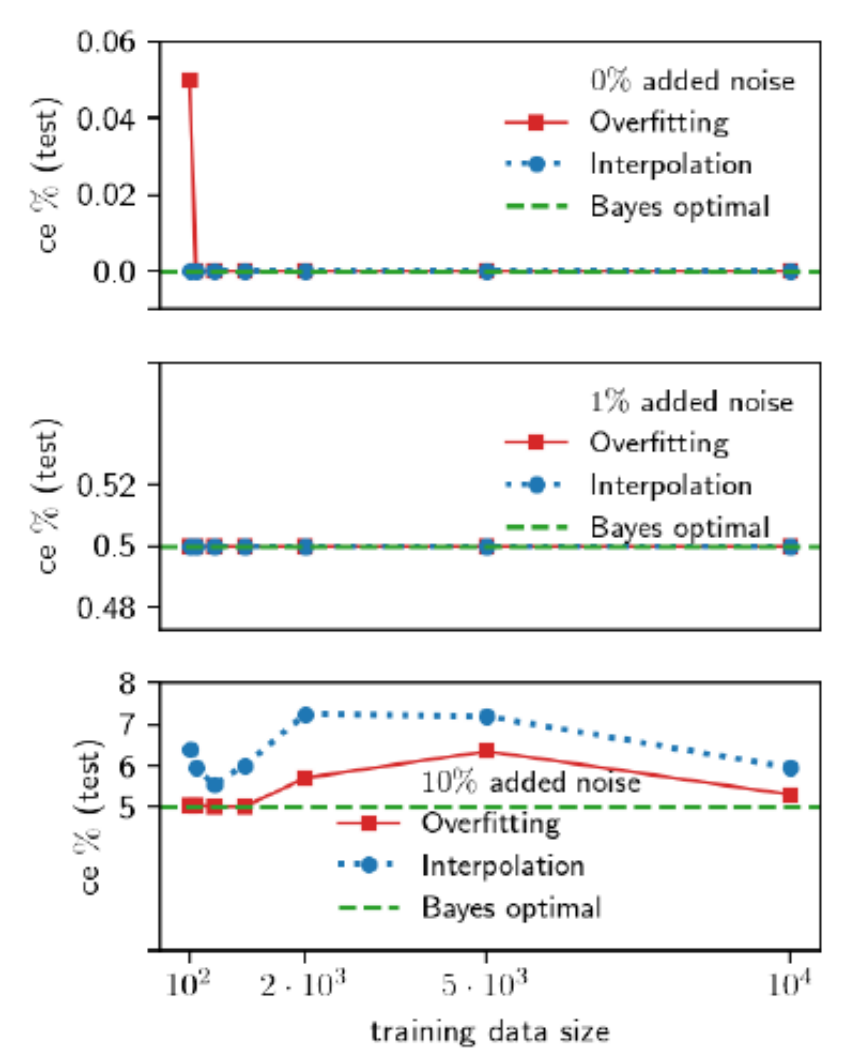
Zero Label Noise?

A potential explanation for disparity in theory and experiments is that there is zero label noise. This can occur in real dataset, e.g., linear separable data.

Experiments Setup

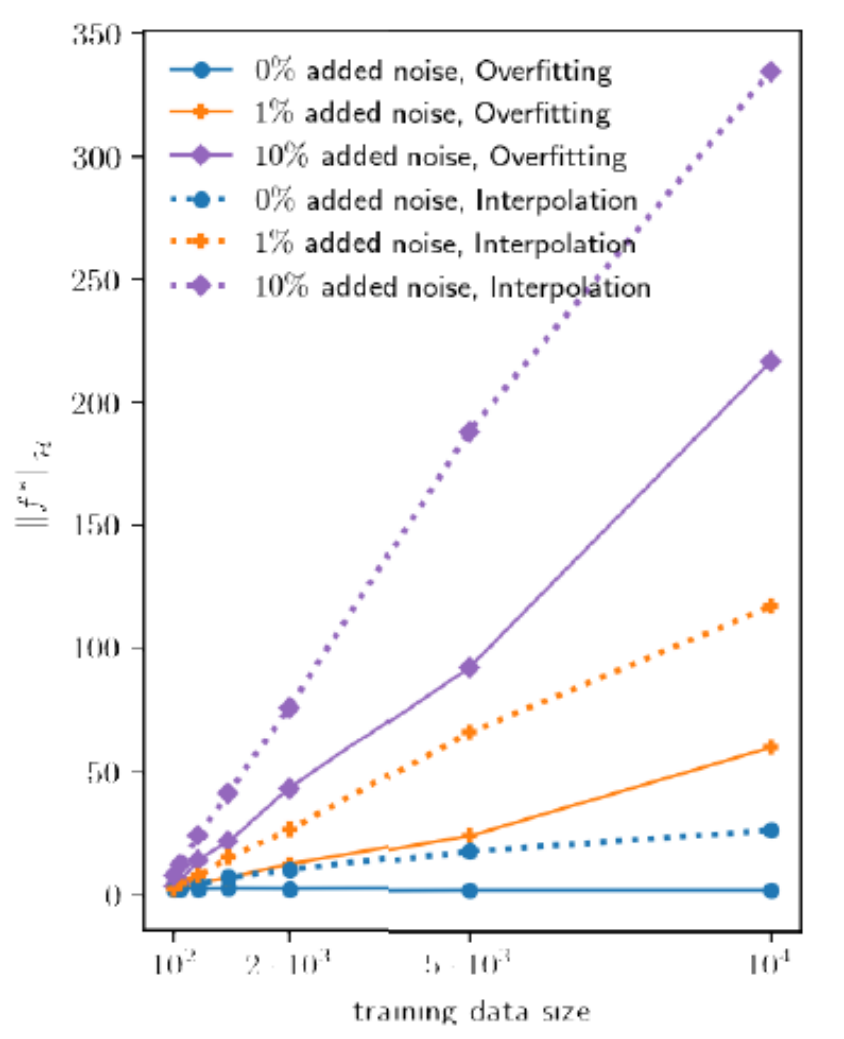
- Dataset: Two synthetic(one separable, one non-separable)
  Real data + noise
- Kernels: Gaussian and Laplacian
- Noise level: 0%, 1%, 10%

Experiment Results(Synthetic Data 1)



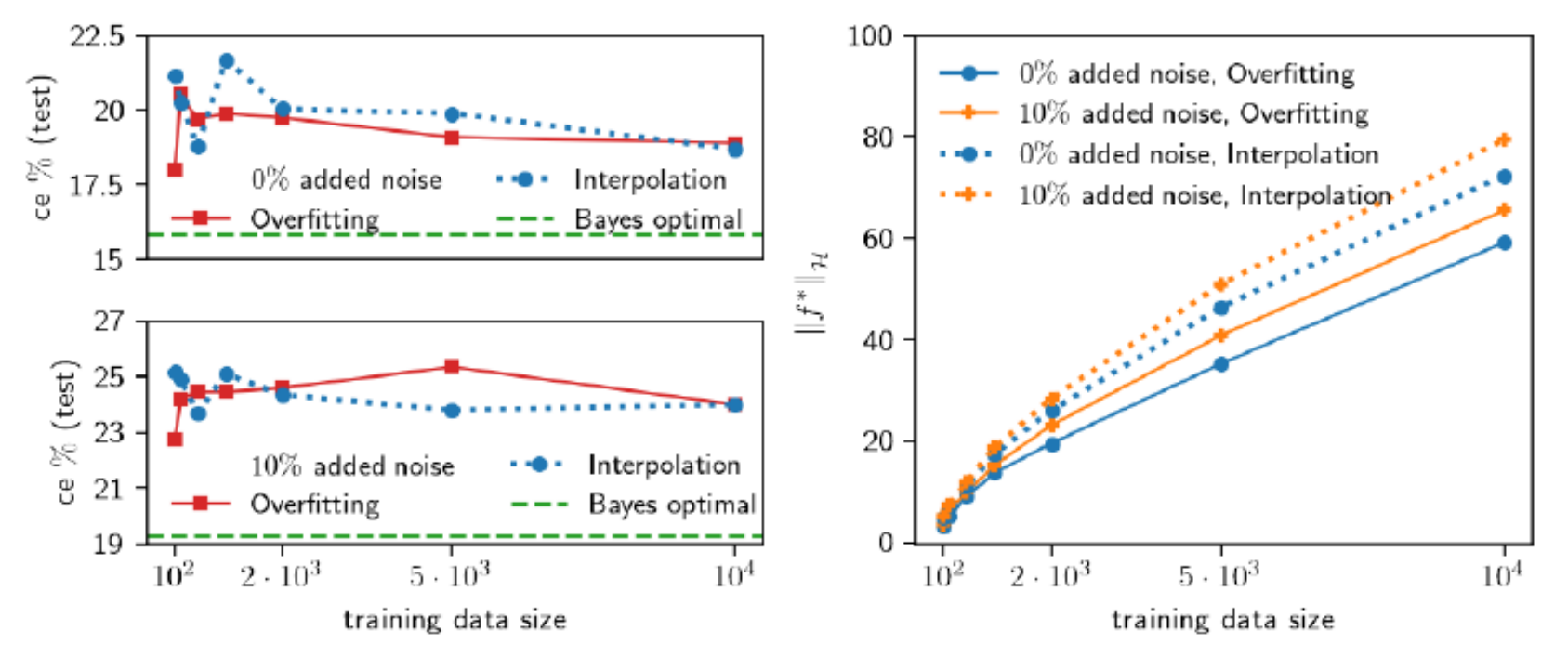
- Overfitting model is the kernel constructed by iterative methods
- Interpolation model is the kernel constructed by direct methods
- Bayes optimal is the label noise
- Both interpolation and overfitting models have close to optimal performance on testing data
- Error rate increases less than label noise

Experiment Results(Synthetic Data 1)



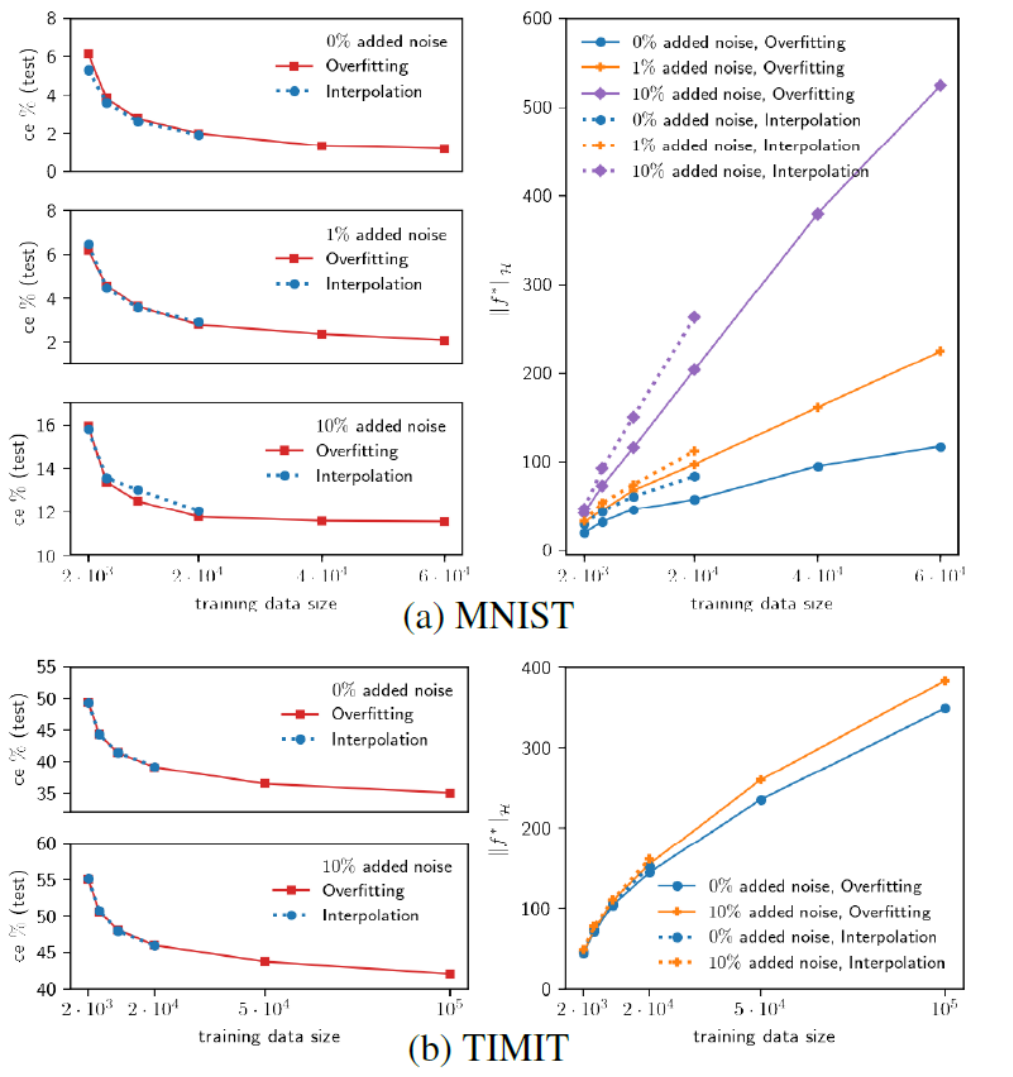
- For linearly separable data, an overfitted solution achieves optimal accuracy with a small norm.
- Adding label noise increases the norm significantly.
- Norm of either solution increases quickly with the number of data points, consistent with our theorem.

Experiment Results(Synthetic Data 2)



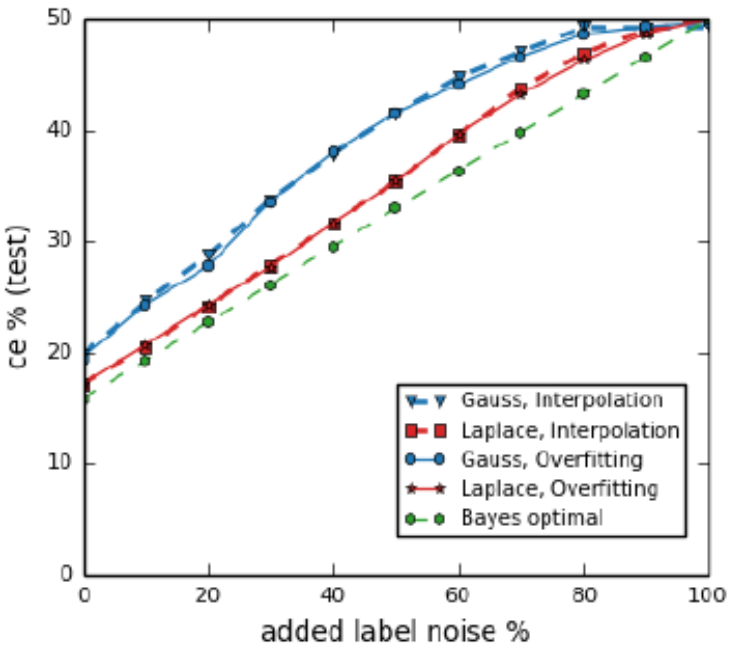
- Both classifiers' performance on test data is within 5% of the Bayes optimal.
- Adding additional label noise should have little impact because the setting is already noisy.

Experiment Results(Real Data + Noise)

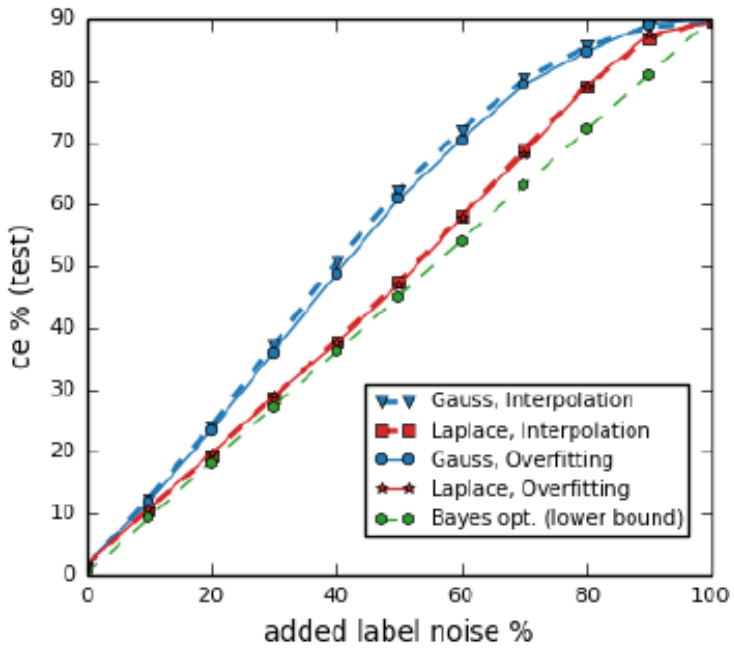


- TIMIT is noisier than MNIST, so the impact of adding label noise is bigger.
- Test performance decays gracefully with amount of label noise.

High Label Noise Bayes Risk Comparison



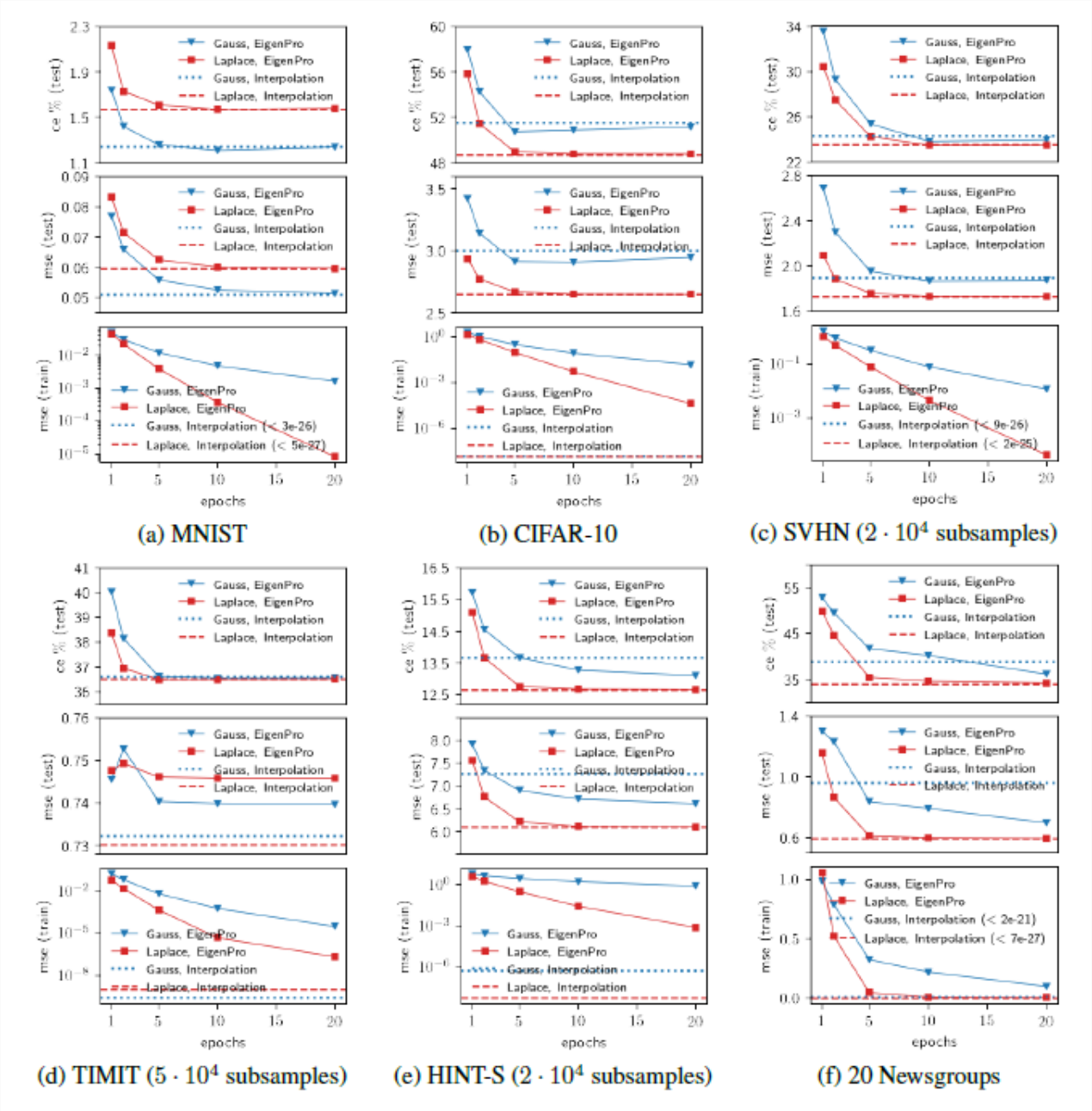
(a) Synthetic-2



(b) MNIST

- All of the classifiers tracks the Bayes risk even for very high level of label noise.
- Laplacian kernel can handle label noise better than Gaussian kernel.

Recap of Experiments Results



(a) MNIST (b) CIFAR-10 (c) SVHN ($2 \cdot 10^4$ subsamples)

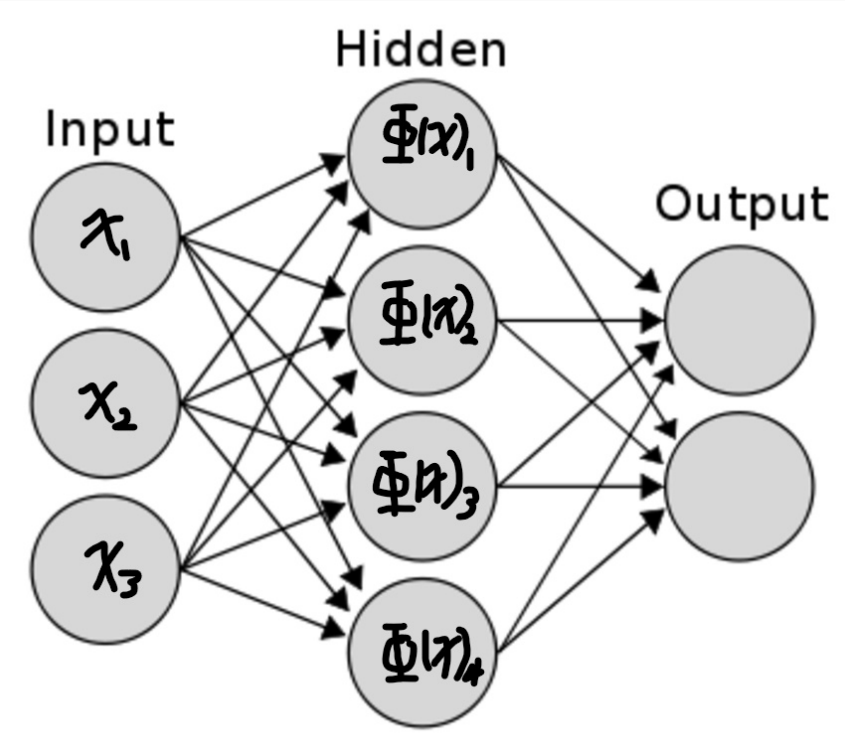
(d) TIMIT ($5 \cdot 10^4$ subsamples) (e) HINT-S ($2 \cdot 10^4$ subsamples) (f) 20 Newsgroups

Laplacian vs Gaussian

- Laplacian kernel takes few iterations to train. Moreover, it can fits random data easily.
- ReLU networks can also fit random data easily[ZBH$^+$16]. This is an interesting similarity between ReLU networks and Laplacian kernel.
- It takes more computational effort to fit a Gaussian kernel.
- Authors of this paper conjecture that this property is related to non-smoothness.
- On the other hand, overfitted/interpolated Gaussian and Laplacian kernels show very similar classification and regression performance on test data and this persists even with added label noise. Hence it appears that the generalization properties of these classifiers are not related to the specifics of the optimization.

Parallels Between Deep and Shallow Architectures in Performance of Overfitted Classifiers

A kernel machine can be viewed as a two layer neural network.



Thus, we can say shallow networks also have nice generalization properties, just like deep ones.

Implicit Regularization and Loss Functions

- Regularization vs Inductive bias: Frequently these two terms are used interchangeably. In this paper, regularization means trading training accuracy for testing performance while inductive bias means gives preferences to certain functions without affecting their output on the training data.

- By this definition, implicit regularization cannot explain the generalization performance, since the training error is 0.

- Another interesting point is that any strictly convex loss function leads to the same interpolated solution. Thus, it is unlikely that the choice of loss function relates to the generalization properties of classifiers.

Wrap-up

- Overfitted/Interpolated kernel classifiers have unexpected good generalization performance, just like deep neural networks.

- Existing theoretical bounds fails to explain this generalization property for kernel classifiers. A close candidate is the bound for 1-nearest neighbour.

Therefore, to understand deep learning, we need to understand kernel learning first!

Questions?

Thank you!

References

[AB09] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[Bel18] Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. *arXiv preprint arXiv:1801.03437*, 2018.

[GK17] Surbhi Goel and Adam Klivans. Eigenvalue decay implies polynomial-time learnability for neural networks. In *Advances in Neural Information Processing Systems*, pages 2192–2202, 2017.

References ii

[RCR15] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[SFB+98] Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

[SSSSC11] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

References iii

[TBRS13] Martin Takác, Avleen Singh Bijral, Peter Richtárik, and Nati Srebro. Mini-batch primal and dual methods for svms. In *ICML (3)*, pages 1022–1030, 2013.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.