The Problem of Generalization

Introduction

1. Underfitting
2. Overfitting

Regularization

1. p=2 ridge regularization / weight decay
2. p=1 lasso regularization

$$J(\theta) = \overbrace{\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f_\theta(x)^{(i)}, y^{(i)})}^{\text{data fit loss}} + \underbrace{\lambda R(\theta)}_{\text{regularizer}} \quad \triangleleft \quad \text{regularized objective function} \tag{11.12}$$

$$R(\theta) = \|\theta\|_p \,. \tag{11.13}$$

The Lp-norm of $\mathbf{x}$ is $(\sum_i |x_i|^p)^{\frac{1}{p}}$. The L_2-norm is the familiar least-squares objective.

Regularizers as Probabilistic Priors

1. "Regularizers can be interpreted as priors that prefer, **a priori** (before looking at the data), some solutions over others."

$$\arg\max_f p\left(f \mid \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N\right) \quad \triangleleft \quad \text{MAP learning} \tag{9.3}$$

2. Using Bayes' rule

$$= \arg\max_f p\left(\{\mathbf{y}^{(i)}\}_{i=1}^N \mid \{\mathbf{x}^{(i)}\}_{i=1}^N, f\right) p(f) \quad \triangleleft \quad \text{by Bayes' rule} \tag{9.4}$$

log posterior is

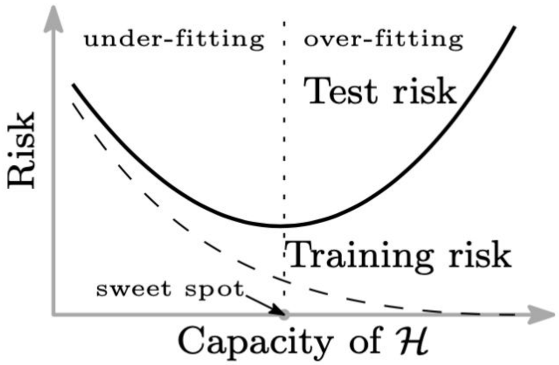
$$J(\theta) = \overbrace{\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_\theta(x)^{(i)}, y^{(i)})}^{\text{data fit loss}} + \underbrace{\lambda R(\theta)}_{\text{regularizer}} \quad \triangleleft \quad \text{regularized objective function} \tag{11.12}$$

Rethinking Generalization

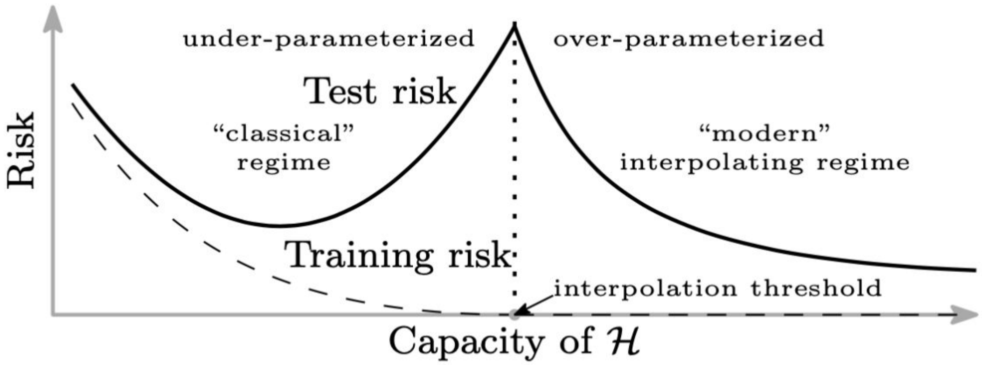
1. Why seemingly complex hypothesis spaces, such as deep nets, tend not to overfit?
2. Some investigations:
    a. state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data
    b. occurs even if we replace the true images by completely unstructured random noise
    c. simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points
    d. The effective capacity of neural networks is sufficient for memorizing the entire data set.
    e. SGD is doing implicit regularization

Rethinking Generalization

1. The double-descent risk curve
2. By considering larger function classes, which contain more candidate predictors compatible with the data, we are able to find interpolating functions that have smaller norm and are thus “simpler”. (Occam’s razor)
3. The number of parameters is just a rough proxy for model capacity



(a)



(b)

Needle in a Haystack

1. Haystack: search space (hypothesis space)
2. Needle: Truth (function that generates our observations)
3. Data: Images/videos
4. Priors: Regularizers (prefer some solutions over others)
5. The hypothesis must be in the hypothesis space
    a. “a drunk man looking for his lost keys under a lamppost . “Why are you looking there,” a cop asks. “Because this is where the light is.”

Needle in a Haystack

1.

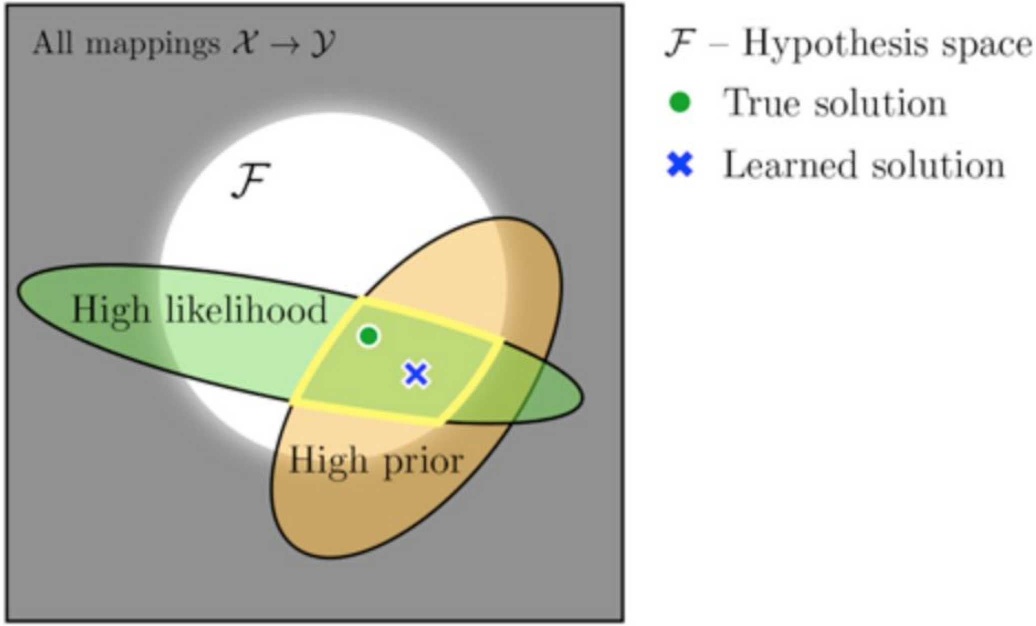


Figure 11.4: A cartoon of the tools for honing in on the truth.

Effect of Data

1. "bottom row, we plot J as a heatmap over the values obtained for different settings of θ.
2. On the top row we plot the data being fit, , along with the function fθ that achieves the best fit, and a sample other settings of θ that achieve within 0.1 of the cost of the best fit."
3. "The more data you have, the less you need other modeling tools"

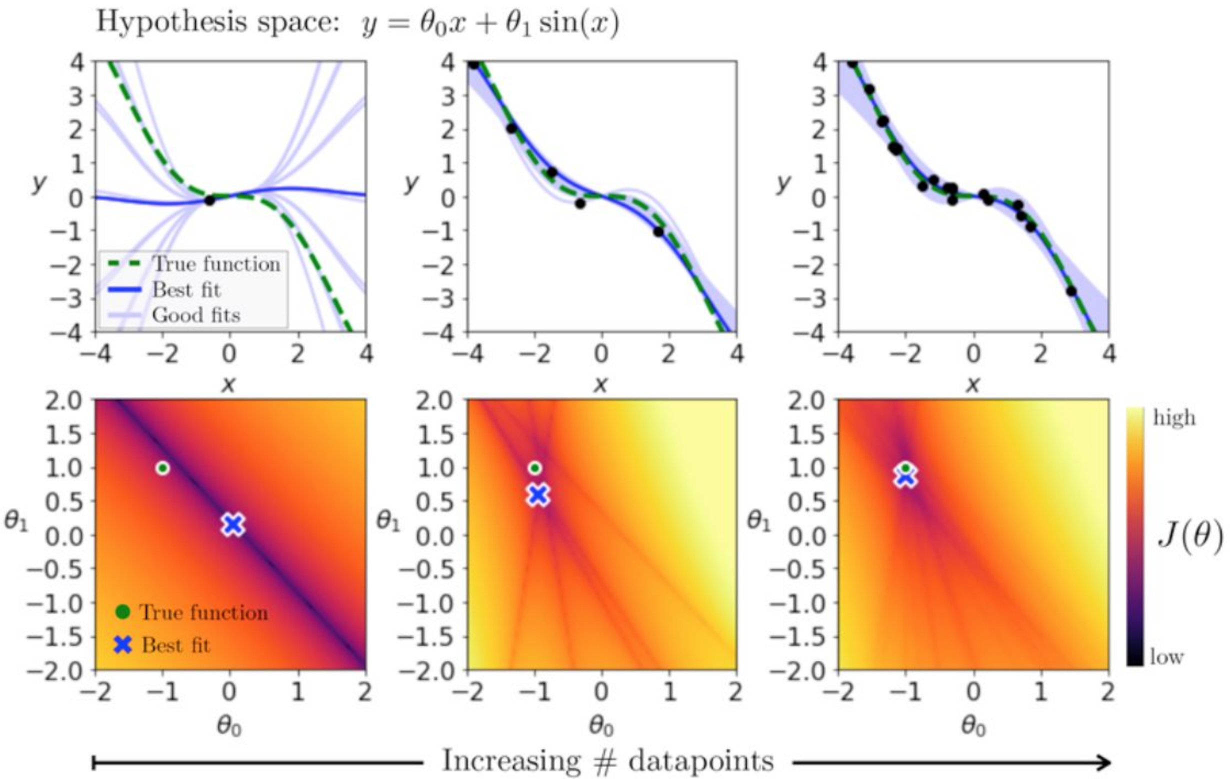


Figure 11.5: More data, more (soft) constraints.

Effect of Priors

$$J(\theta; \{x^{(i)}, y^{(i)}\}_{i=1}^N) = \frac{1}{N}\sum_i \|f_\theta(x^{(i)}) - y^{(i)}\|_2^2 + \lambda \|\theta\|_2^2 \quad \triangleleft \text{ objective} \tag{11.16}$$

$$f_\theta(x) = \theta_0 x + \theta_1 x \quad \triangleleft \text{ hypothesis space} \tag{11.17}$$

1. “Priors help only when they are good guesses as to the truth.”
2. “Over Reliance on the prior means ignoring the data, and this is generally a bad thing.”
3. “For any given prior, there is a sweet spot where the strength is optimal.”

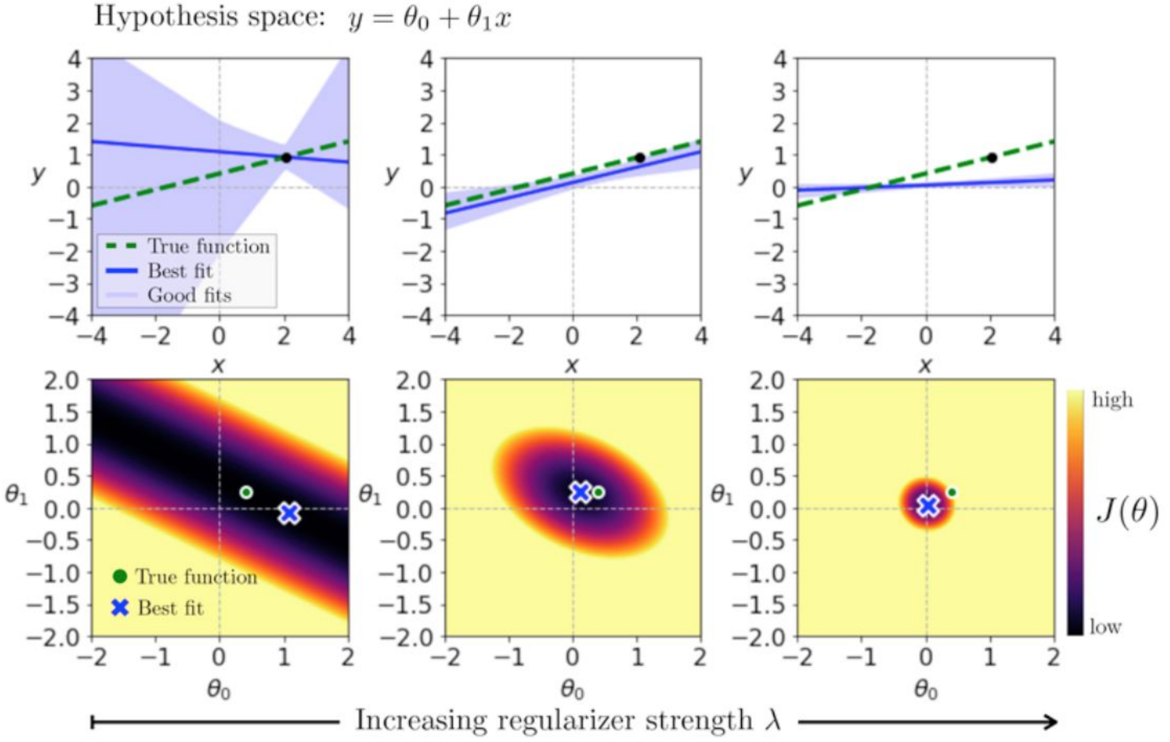


Figure 11.6: More regularization, more (soft) constraints.

Effect of Hypothesis Space

$$f_\theta(x) = \theta_0 x + \theta_1 x^2 \quad \triangleleft \ \texttt{quadratic} \quad (11.18)$$

$$f_\theta(x) = \theta_0 x \quad \triangleleft \ \texttt{linear} \quad (11.19)$$

$$f_\theta(x) = 0 \quad \triangleleft \ \texttt{constant} \quad (11.20)$$

1. "Using a smaller hypothesis space can potentially accelerate our search"

2. But don't go too far!

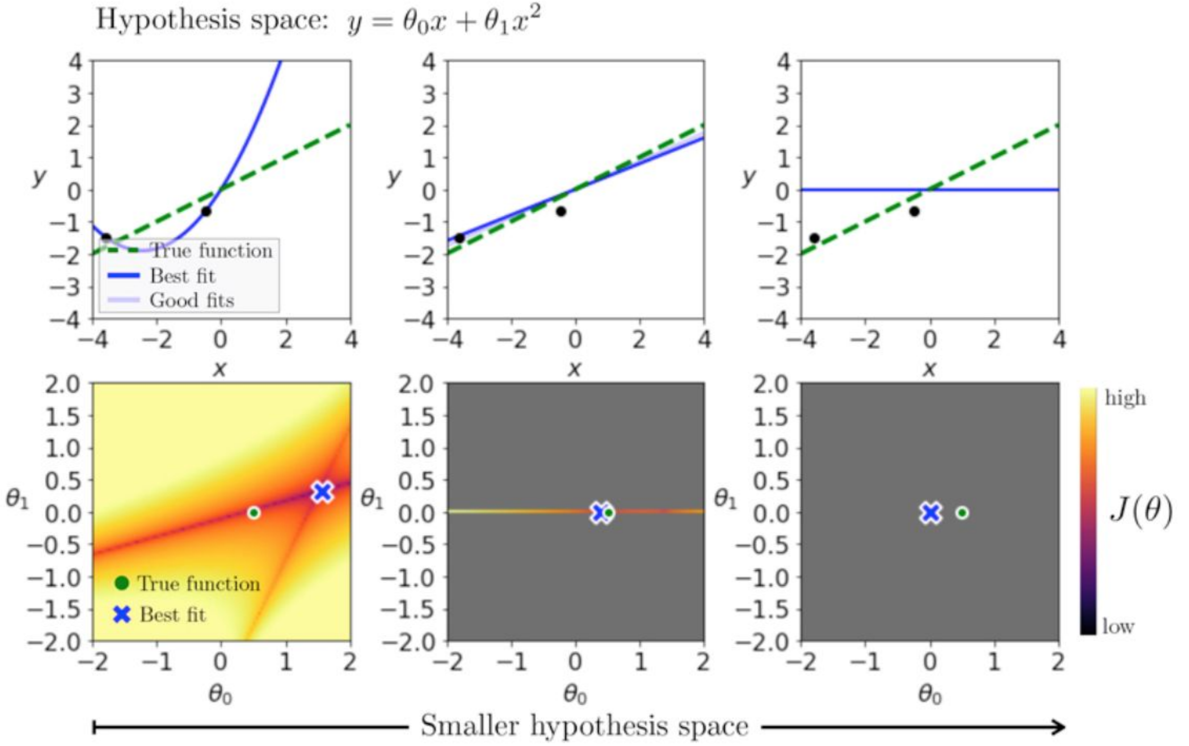


Figure 11.7: Fewer hypotheses, more (hard) constraints.

Takeaway

1. If you don't have much data,
    a. you can use strong priors and
    b. structural constraints instead.
2. If you don't have much domain knowledge,
    a. you can collect lots of data instead

Model-based vs Learning-based

1. “One goal of learning algorithms is to make systems that generalize ever better, meaning they continue to work even when the test data is very different than the training data.”
2. “Currently, however, the systems that generalize in the strongest sense—that work for all possible test data—are generally not learned but designed according to other principles.”
3. “In this way, many classical algorithms still have advantages over the latest learned systems. But this gap is rapidly closing!”

References

1. Foundations of Computer Vision - Chapter 11
2. [1611.03530] Understanding deep learning requires rethinking generalization
3. [1812.11118] Reconciling modern machine learning practice and the bias-variance trade-off