

6.962 Week 4

Topic: Statistical Learning Theory

Presenter: Emin Martinian

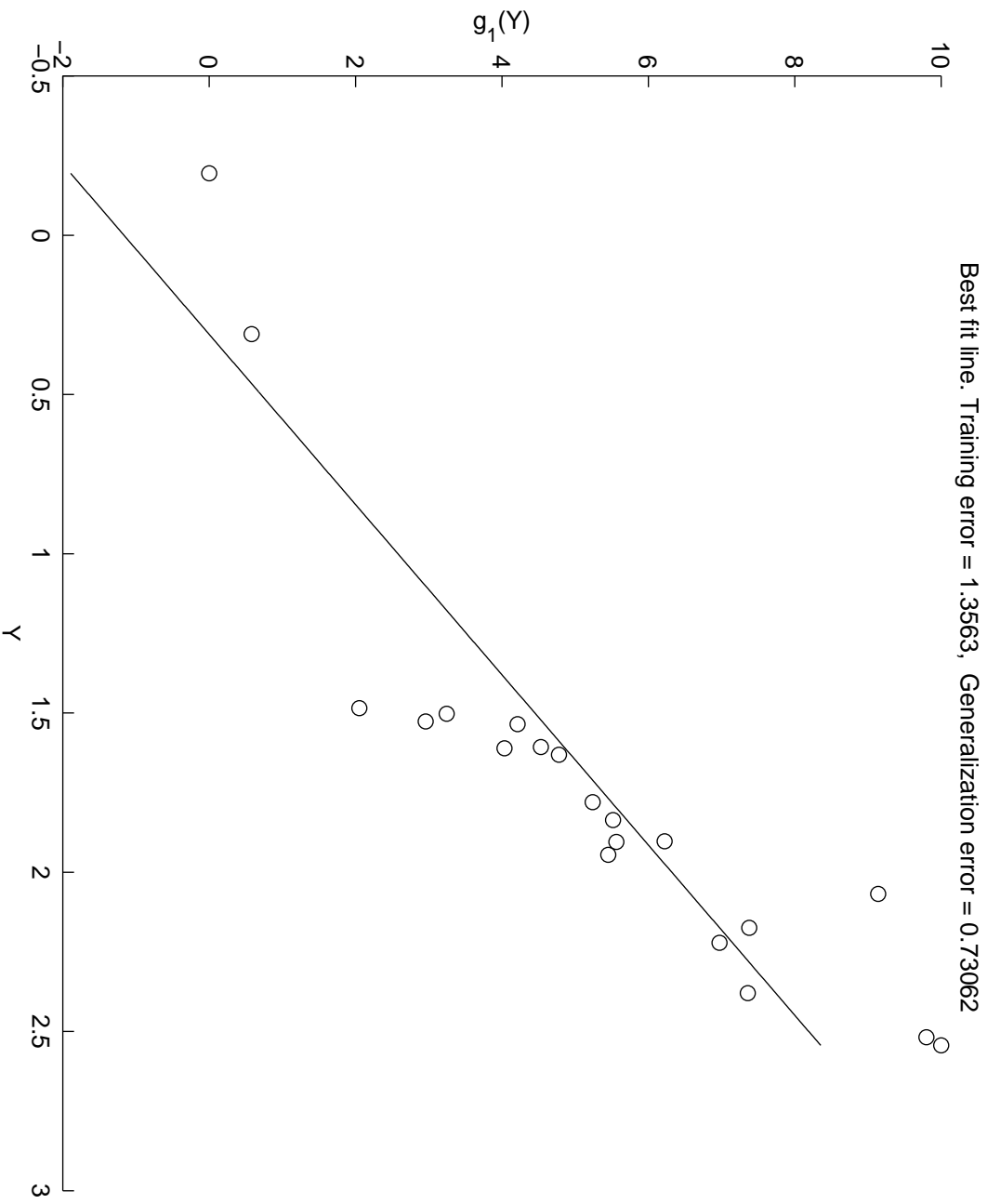
- Random variables X and Y with unknown density $p(x, y)$.
- Observe Y and estimate $\hat{X} = g(Y)$.
- *M* i.i.d. training samples (x_i, y_i) are available.

Estimation Structure:

- Parametric set of estimators $g(Y, \alpha)$, $\alpha \in \Lambda$.
- Λ is an arbitrary parameter set.
- Since Λ is arbitrary, no loss of generality.
- Goal: use training samples (x_i, y_i) to find good parameter α^* .

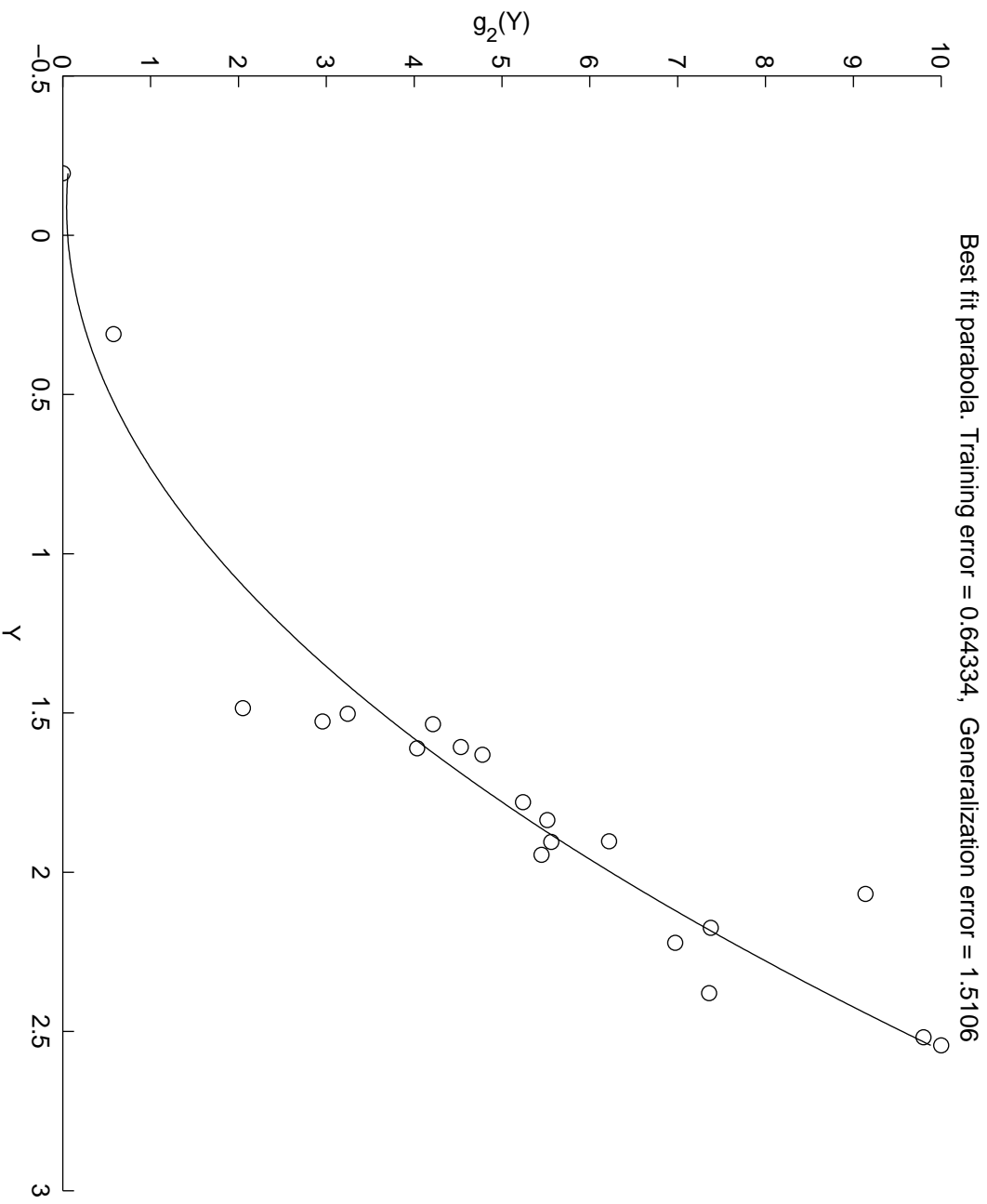
Simple Example: Least square line fit

- $\Lambda = \mathbf{R}^2$, $g(Y, \alpha_0, \alpha_1) = \alpha_0 + \alpha_1 \cdot Y$.
- How do we choose α ?
- ERM Principle \Rightarrow choose α which minimizes training error.

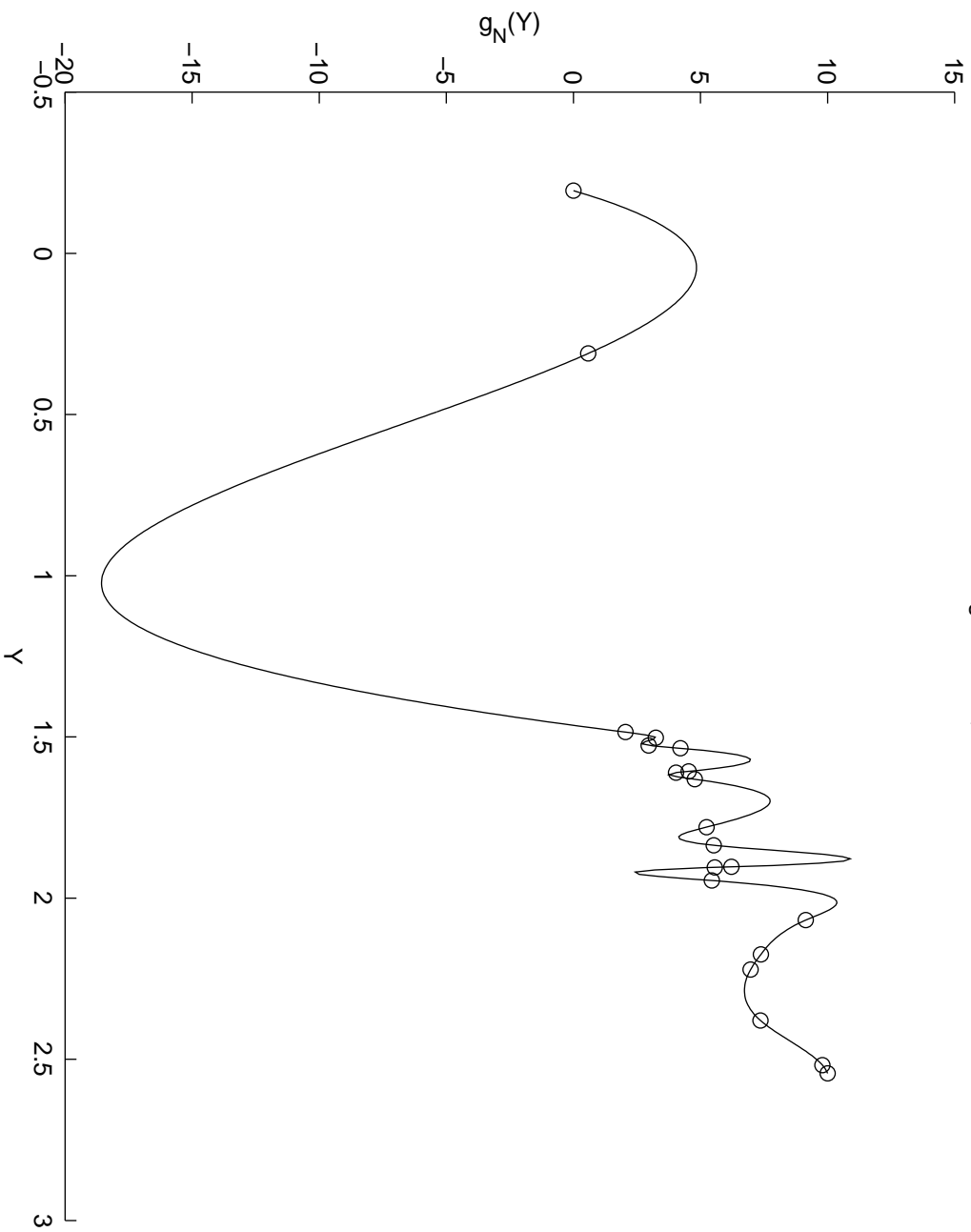


- $M = 20$ data points generated according to
$$Y = \log(1 + X) + N.$$
- X uniform over $[1, 10]$, N is 0-mean Gaussian, $\sigma = .15$.
- Training error = 1.37, Generalization error = .73
- Can we do better?

- We can use models richer than simple line.
- e.g. higher order polynomials, splines, neural networks, etc.
- How do we choose α for more complicated models?

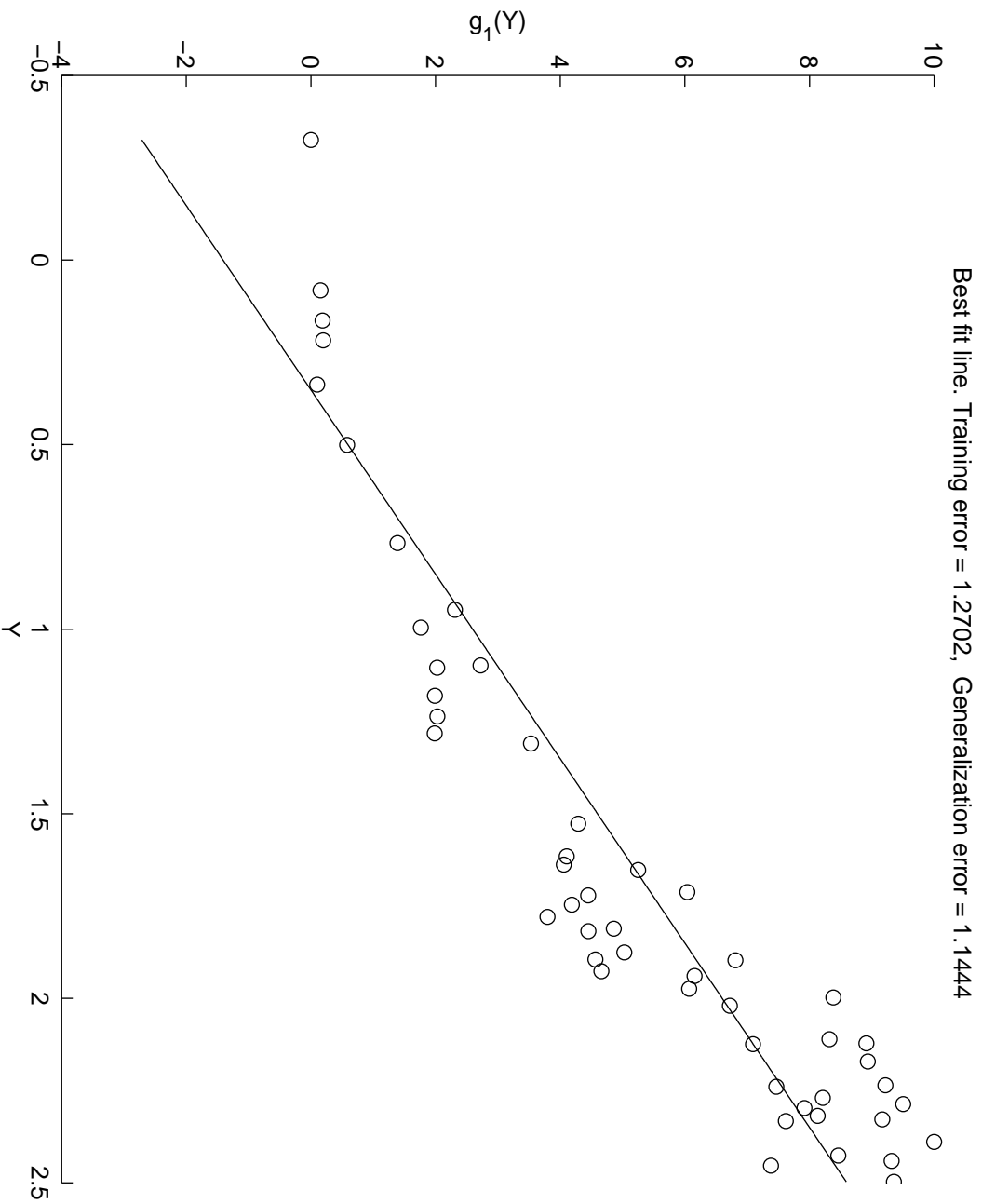


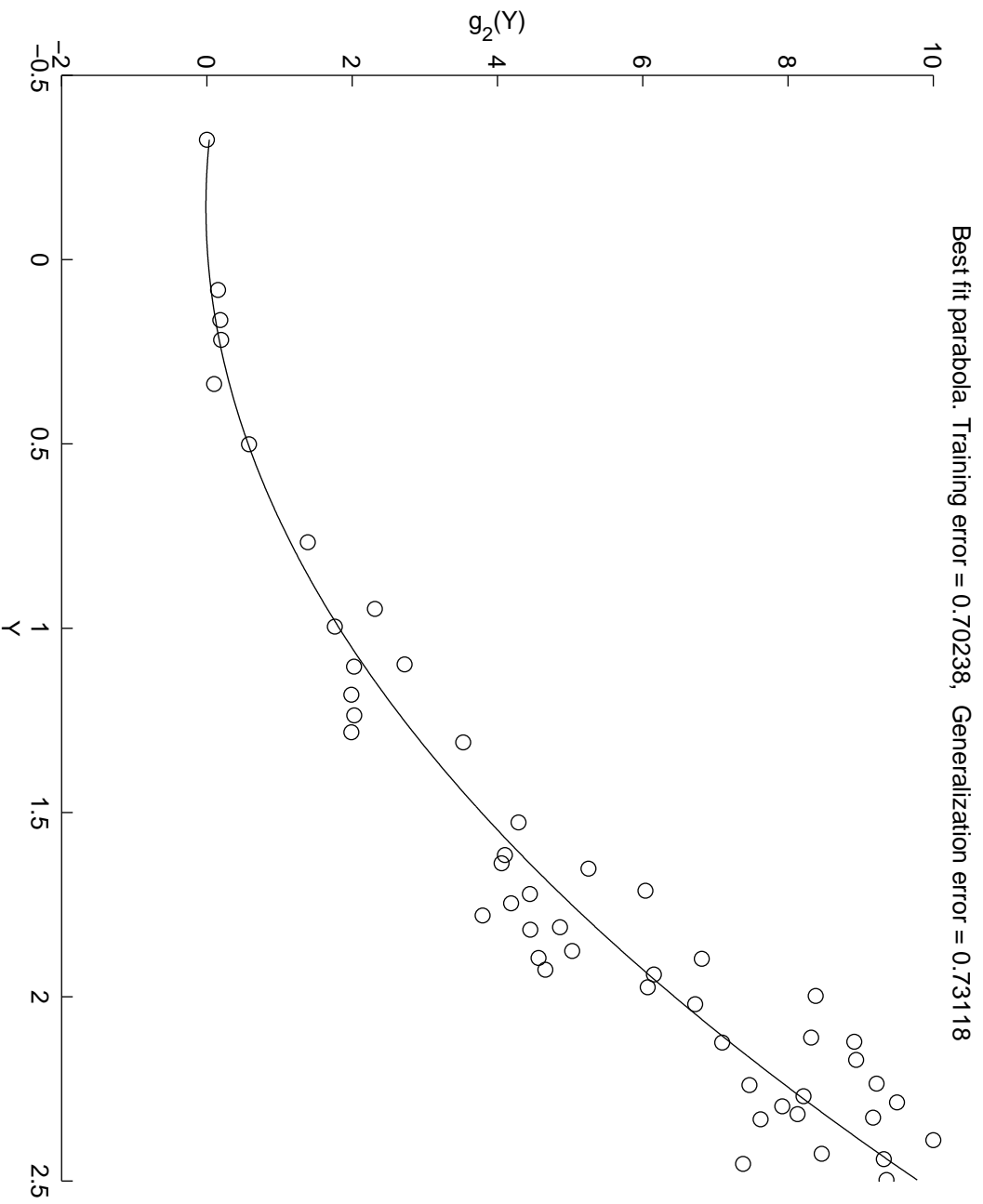
Best fit Mth order curve. Training error = 0, Generalization error = 80.886



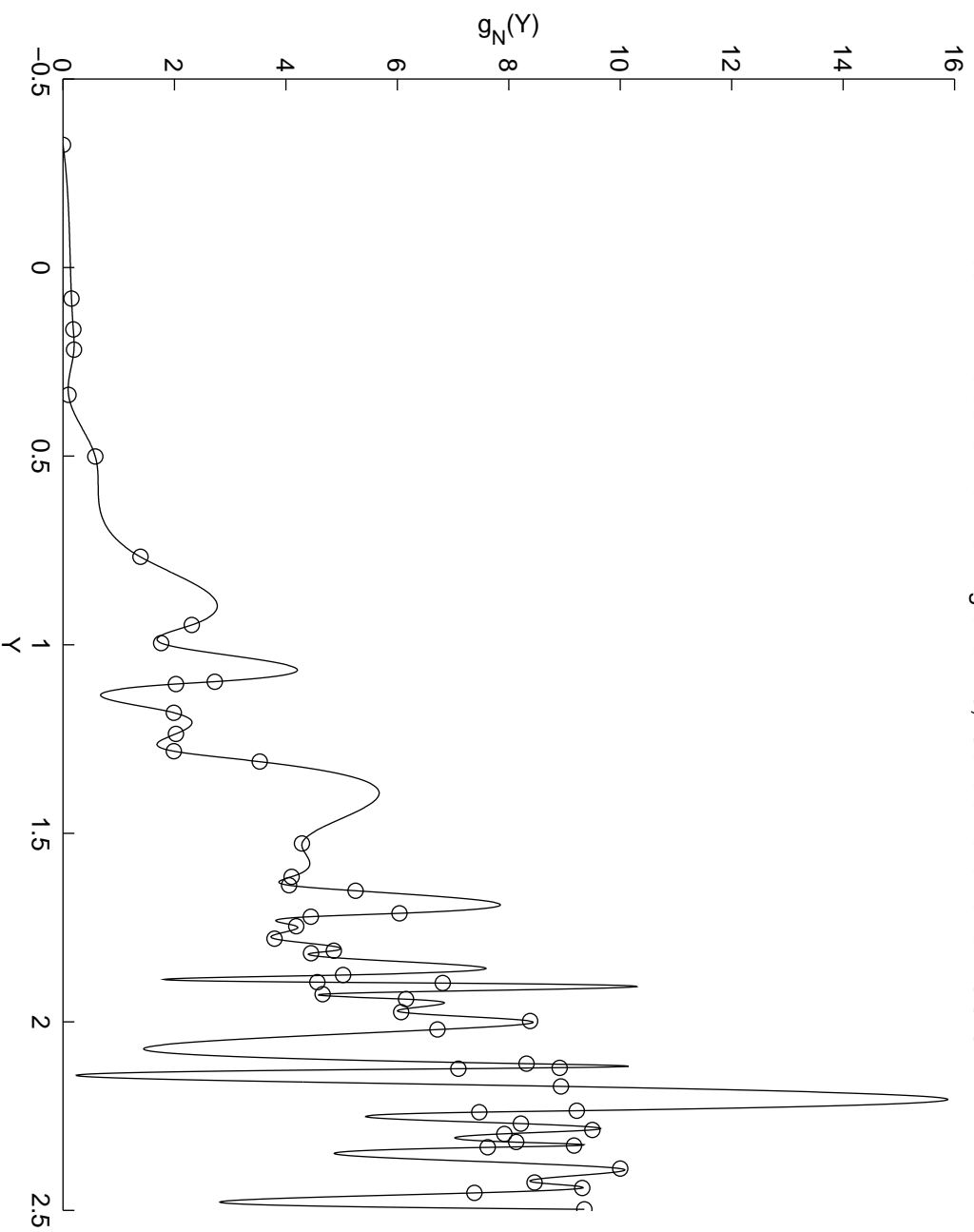
Curve order	Training Error	Generalization Error
1	1.37	0.73
2	0.64	1.51
$M = 20$	0	80.89

- Richer models lower training error, raise generalization error!
- Choosing α harder for complicated models (need more data).





Best fit Mth order curve. Training error = 0, Generalization error = 5.3282



Curve order	Training Error	Generalization Error
1	1.27	1.14
2	0.70	0.73
$M = 50$	0	5.33

- Enough data for good parabolic fit.
- M th order curve still too complicated to find good α .
- How do we know when we have enough data?

Observations:

- With few training samples, simple estimator better.
- With more samples, richer estimator becomes better.
- Very rich models explain data perfectly, but don't generalize.

- Observations suggest the idea of capacity control:
 - With sparse data use simple models.
 - With more data use richer models.
 - Ockam's razor, Minimum Description Length Principle, etc.
- These are all heuristic, can we get rigorous version?

Simplest Model:

- $\mathcal{X}, \mathcal{Y}, \Lambda$ are finite sets.
- Loss function is probability of error $P(\alpha_l)$.
- Weak LLN (Chebyshev version) says

$$\Pr[|V(\alpha_l) - P(\alpha_l)| > \epsilon] \leq \frac{\epsilon^2}{\sigma_l^2}$$

- Provides way to determine how much data required, but generally requires knowledge of probability distribution.

- Chernoff Bound: $\Pr[|V(\alpha_l) - P(\alpha_l)| > \epsilon] \leq 2e^{-2M\epsilon^2}$
- Let $E_i(\alpha_l)$ indicate whether $g(Y_i, \alpha_l) = X_i$.
- $\xi_l = V(\alpha_l) - P(\alpha_l) = \frac{1}{M} \sum_{i=1}^M E_i(\alpha_l) - P(\alpha_l)$
- $\xi_l \rightarrow \mathcal{N}(0, \sqrt{P(\alpha_l)(1 - P(\alpha_l))}) / \sqrt{M}$ according to CLT.
- $\Pr[|\xi_l| > \epsilon] \leq 2\mathcal{Q}\left(\epsilon \sqrt{\frac{M}{P(\alpha_l)(1 - P(\alpha_l))}}\right) \leq 2\mathcal{Q}\left(\epsilon \sqrt{\frac{M}{(1/4)}}\right)$
- Using $\mathcal{Q}(x) \leq e^{-\frac{x^2}{2}}$ we get $\Pr[|\xi_l| > \epsilon] \leq 2e^{-2M\epsilon^2}$.

- $\Pr[|V(\alpha_l) - P(\alpha_l)| > \epsilon] \leq 2e^{-2M\epsilon^2}$ bounds probability that $g(\cdot, \alpha_l)$ does well on training data but poorly in the future.
- What we really want to bound is the probability that the chosen estimator, $g(\cdot, \alpha^*)$ does poorly in the future.

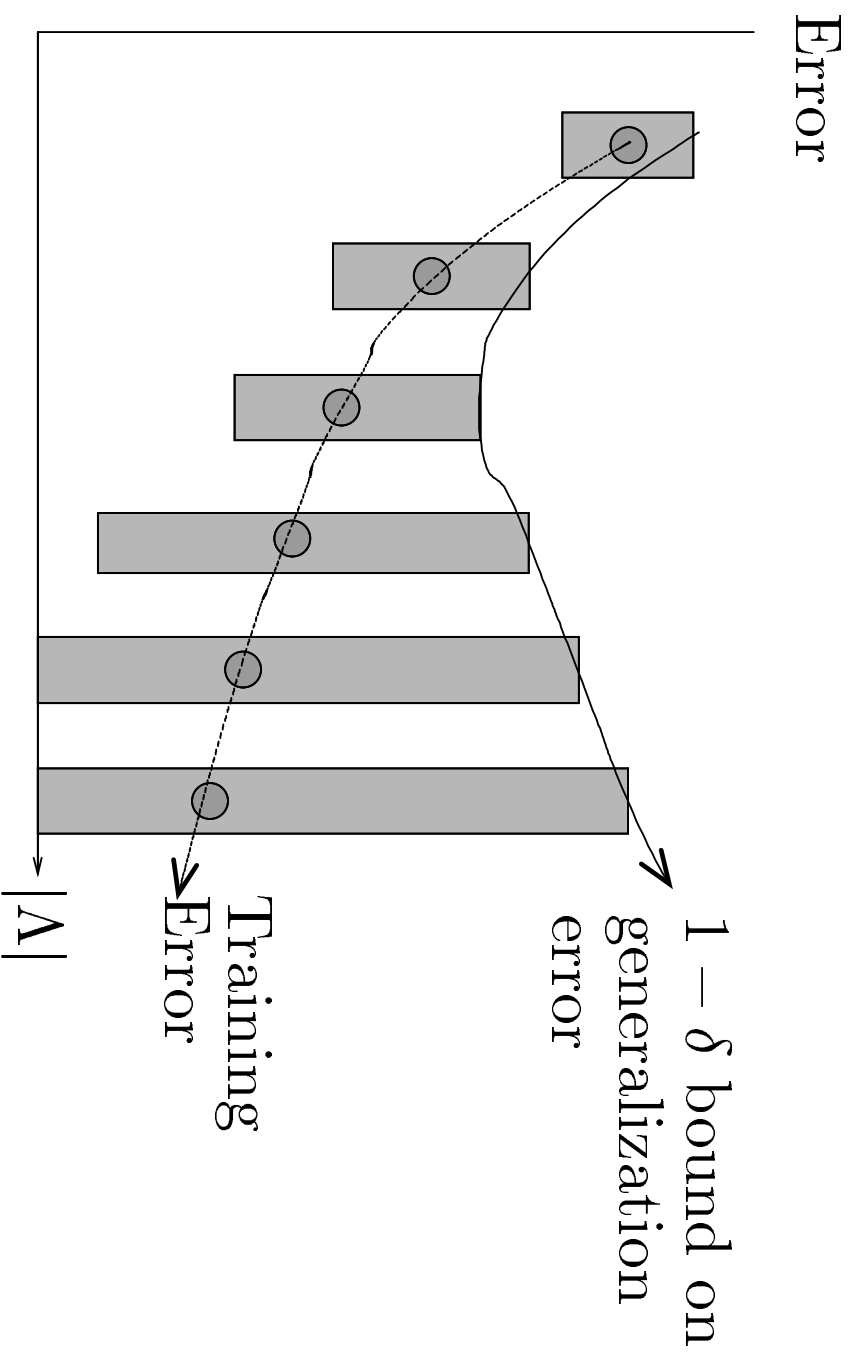
$$\begin{aligned} \Pr[|V(\alpha^*) - P(\alpha^*)| > \epsilon] &\leq \Pr\left[\sup_{\alpha \in \Lambda} |V(\alpha) - P(\alpha)| > \epsilon\right] \\ &\leq \sum_{l=1}^M \Pr[|V(\alpha_l) - P(\alpha_l)| > \epsilon] \\ &\leq 2|\Lambda|e^{-2M\epsilon^2} \end{aligned}$$

$$\Pr[|V(\alpha^*) - P(\alpha^*)| > \epsilon] \leq 2 \exp \left[-M \left(2\epsilon^2 - \frac{\log |\Lambda|}{M} \right) \right] = \delta$$

- $1 - \delta$ confidence interval: $\epsilon \pm \sqrt{\frac{1}{2M} \log \frac{|\Lambda|}{\delta/2}}$
- Trade-off between accuracy and generalization.
- Better generalization requires lowering $|\Lambda|$ or raising M .

Structural Risk Minimization at confidence $1 - \delta$:

- Choose structure of estimation models
 - $A_1 \subset A_2 \subset \dots \subset A_n \subset \dots$
 - $|A_1| < |A_2| < \dots < |A_n| < \dots$
- Choose $\alpha_n^* \in A_n$ to minimize $V(\alpha_n^*) + \sqrt{\frac{1}{2M} \log \frac{|A_n|}{\delta/2}}$.
- Asymptotically, SRM and FERM choose α^* to minimize $V(\alpha_n^*)$.



- What if \mathcal{X} , \mathcal{Y} , Λ not finite?
- For M training samples (x_i, y_i) each estimator $g(\cdot, \alpha)$ can be characterized by error vector $\vec{E}(\alpha)$.
- $E_i(\alpha) = 1$ if and only if $g(\cdot, \alpha)$ wrong on (x_i, y_i) .
- Estimators with same error vectors indistinguishable. 2^M possible error vectors \Rightarrow at most 2^M distinguishable estimators.

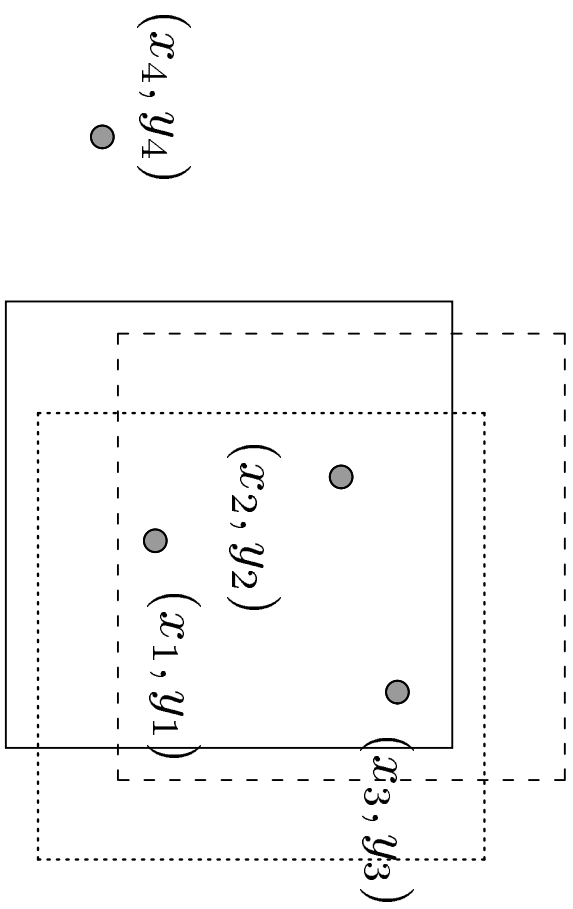


Figure 1: Squares showing error indicator functions for estimators $g(\cdot, \alpha_1)$, $g(\cdot, \alpha_2)$, and $g(\cdot, \alpha_3)$. $E(\alpha_1) = E(\alpha_2) = E(\alpha_3) = (1, 1, 1, 0)$.

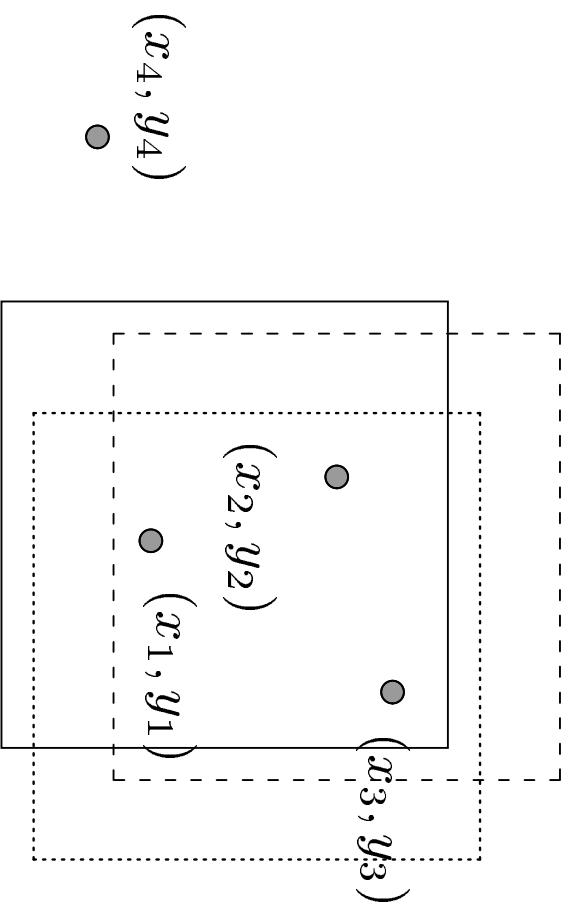


Figure 2: If the error indicator functions in Λ all correspond to squares, then the error vector $E(\alpha) = (1, 0, 0, 1)$ is impossible for this set of points. In fact for any set of 4 points it is impossible to get all 16 error vectors.

- The log of the maximum number of distinguishable estimator–error vector pairs in Λ is called the growth function $G^\Lambda(M)$.
- If $G^\Lambda(M) = M \log 2$ for all M then our bounds won't work.
- Previous example shows $G^\Lambda(M) < M \log 2$ for $M = 4$ so we can try to use bounds even though \mathcal{X} , \mathcal{Y} , and Λ are not finite.

- Using some clever symmetry arguments we can show

$$\Pr \left[\sup_{\alpha \in \Lambda} |V(\alpha) - P(\alpha)| > \epsilon \right] \leq 6 \exp \left[-M \frac{\epsilon^2}{4} + G^\Lambda(2M) \right]$$

- Similar to previous bound for finite Λ .
- Calculating $G^\Lambda(M)$ for all M seems difficult.

- Vapnik showed only 2 possible forms for $G^\Delta(M)$:
 - $G^\Delta(M) = M \log 2$.
 - $\exists h$ such that $\forall M > h, G^\Delta(M) \leq h(1 + \log \frac{M}{h})$.
- If h exists it is called the VC-dimension.
- h is the largest finite set which can be shattered by A .

Using VC-dimension, h to bound M we obtain the following theorem:

If $\alpha \in \Lambda$ indexes a set of indicator functions with corresponding relative frequencies $V(\alpha)$, then for a sequence of M i.i.d. random variables with common distribution P

$$\Pr \left\{ \sup_{\alpha \in \Lambda} |P(\alpha) - V(\alpha)| > \epsilon \right\} < 4 \exp \left\{ -M \left[(\epsilon - M^{-1})^2 - \frac{h(1 + \log(2M/h))}{M} \right] \right\}.$$

- The previous bound can be used for non-finite Λ as long as the VC-dimension, h can be computed.
- It turns out that h is the supremum of the largest finite set that can be shattered by an element of Λ .
- Previous results regarding Structural Risk Minimization can be extended to non-finite Λ and more general error measures.

Conclusions:

- Given: M training samples (x_i, y_i) from i.i.d. distribution $p(x, y)$.
- Goal: find best estimator $g(Y, \alpha^*)$, $\alpha^* \in A$.
- Result: Minimizing training error (empirical risk) is not always good. Minimizing structural risk via capacity control better.
- Tools: VC-dimension formulas give non-asymptotic (i.e. valid for all M), bounds on successful generalization.