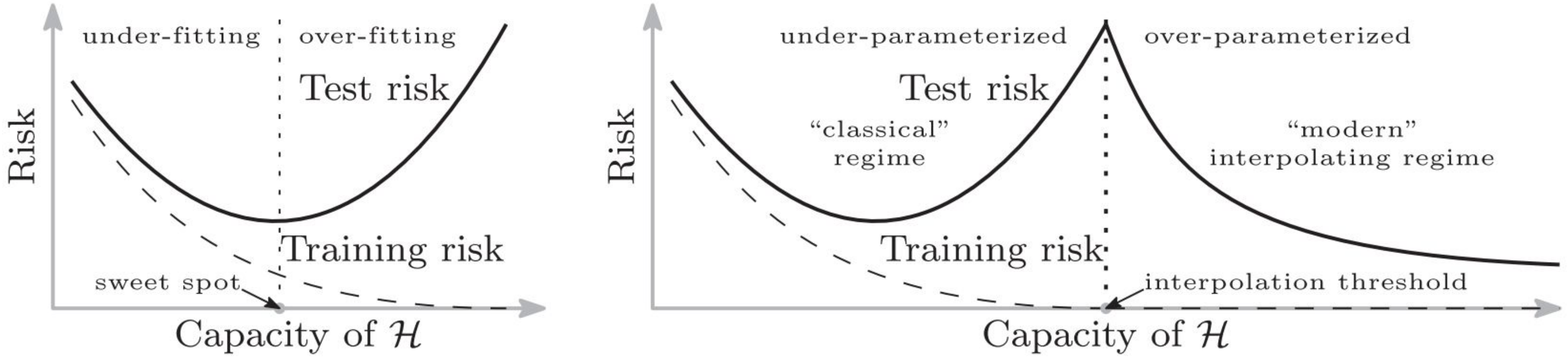
Double Descent

Alexey Voskoboinikov

Double Descent

Classical generalization curve

Modern generalization curve



Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

Linear model

$$h(x) = \sum_{k=1}^{N} \theta_k \phi(x, v_k)$$

Random Fourier features:

1. $\phi(x, v_k) = e^{i\langle x, v_k \rangle}, \quad v_k \sim \mathcal{N}(v_k | 0, \alpha I)$

2. $\phi(x, v_k) = \sqrt{2} \cos(\langle x, v_k \rangle + b_k)$

$v_k \sim \mathcal{N}(v_k | 0, \alpha I), \quad b_k \sim \mathbf{Unif}(0, 2\pi)$

Random ReLU features:

3. $\phi(x, v_k) = \max(0, \langle x, v_k \rangle)$

$v_k \sim \mathcal{N}(v_k | 0, \alpha I)$

Underparameterized regime

Excess risk:

$$R(\hat{\theta}) = ||\hat{\theta} - \theta^*||_2^2$$

Expected excess risk:

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2 \mathbb{E}\left[\mathbf{tr}((\Phi^T \Phi)^{-1})\right] = \sigma^2 \frac{d}{n-d-1}$$

Overparameterized regime

Excess risk:

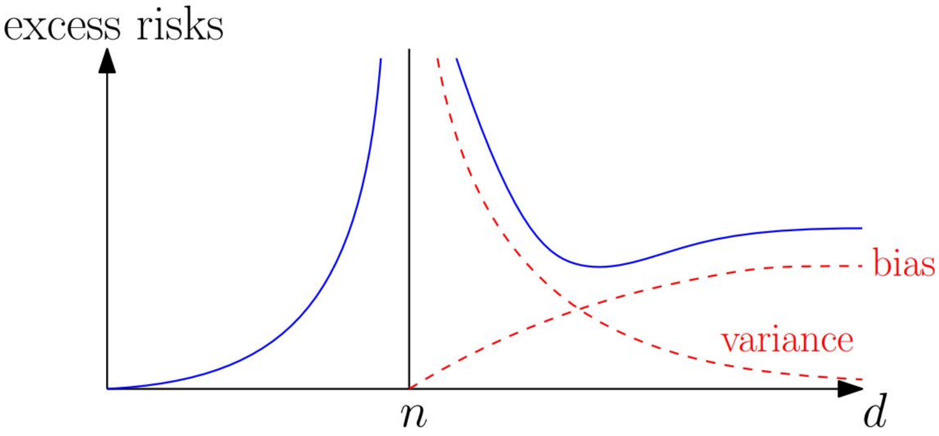
$$R(\hat{\theta}) = ||\hat{\theta} - \theta^*||_2^2$$

Expected excess risk:

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2 \frac{n}{d-n-1} + ||\theta^*||_2^2 \frac{d-n}{d}$$

Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

Linear model



Underparameterized regime

Excess risk:

$$R(\hat{\theta}) = ||\hat{\theta} - \theta^*||_2^2$$

Expected excess risk:

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2 \mathbb{E}\left[\mathbf{tr}((\Phi^T\Phi)^{-1})\right] = \sigma^2 \frac{d}{n-d-1}$$

Overparameterized regime

Excess risk:

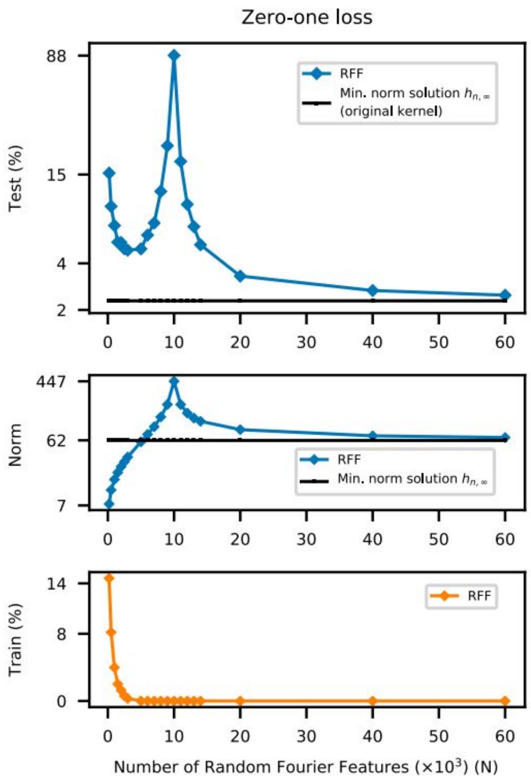
$$R(\hat{\theta}) = ||\hat{\theta} - \theta^*||_2^2$$

Expected excess risk:

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2 \frac{n}{d-n-1} + ||\theta^*||_2^2 \frac{d-n}{d}$$

Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

Linear model



Underparameterized regime

Excess risk:

$$R(\hat{\theta}) = ||\hat{\theta} - \theta^*||_2^2$$

Expected excess risk:

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2 \mathbb{E}\left[\mathbf{tr}((\Phi^T\Phi)^{-1})\right] = \sigma^2 \frac{d}{n-d-1}$$

Overparameterized regime

Excess risk:

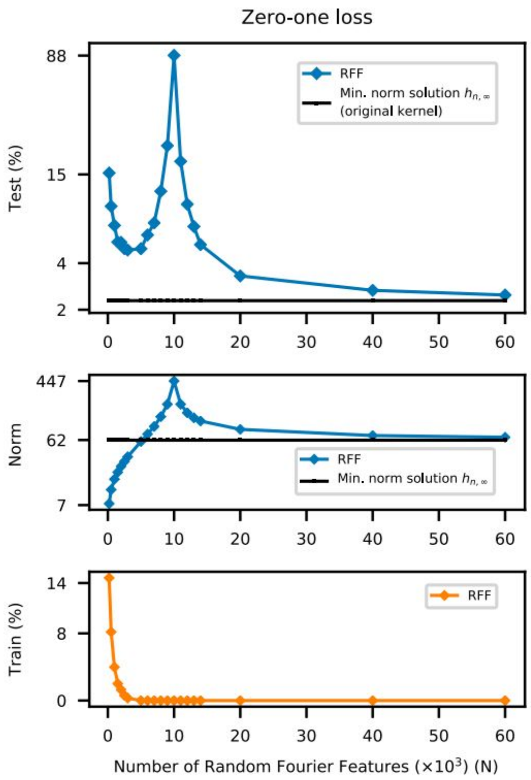
$$R(\hat{\theta}) = ||\hat{\theta} - \theta^*||_2^2$$

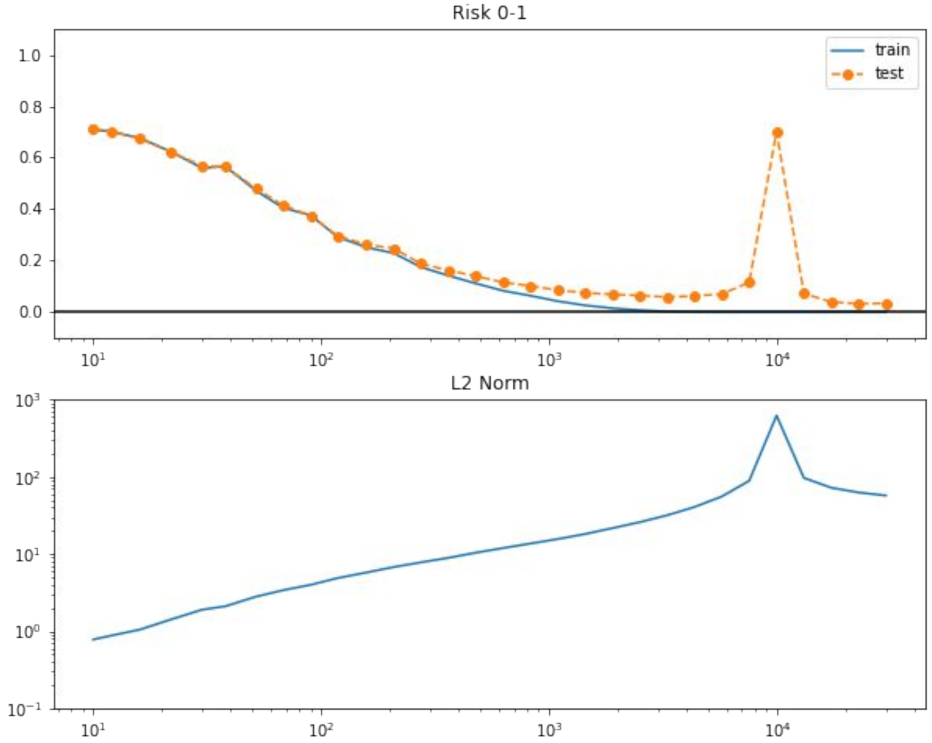
Expected excess risk:

$$\mathbb{E}\left[R(\hat{\theta})\right] = \sigma^2 \frac{n}{d-n-1} + ||\theta^*||_2^2 \frac{d-n}{d}$$

Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

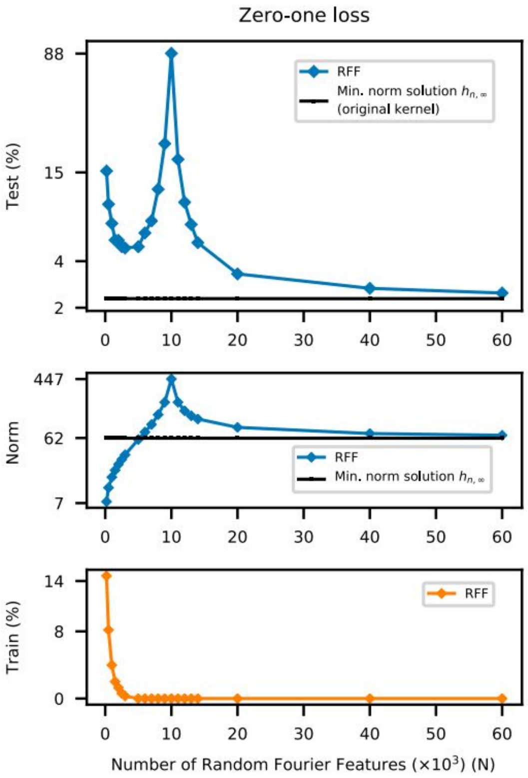
Linear model



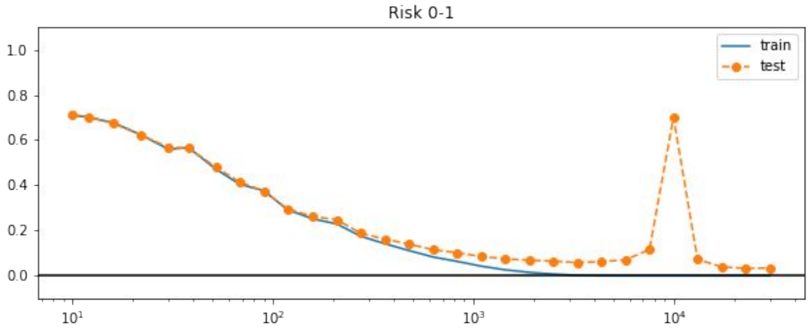


Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

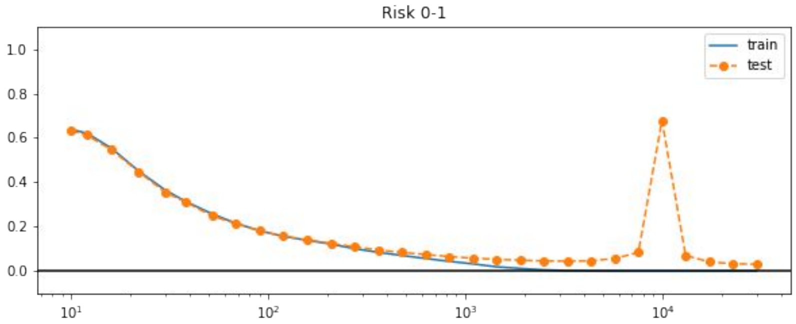
Linear model



Random Fourier features

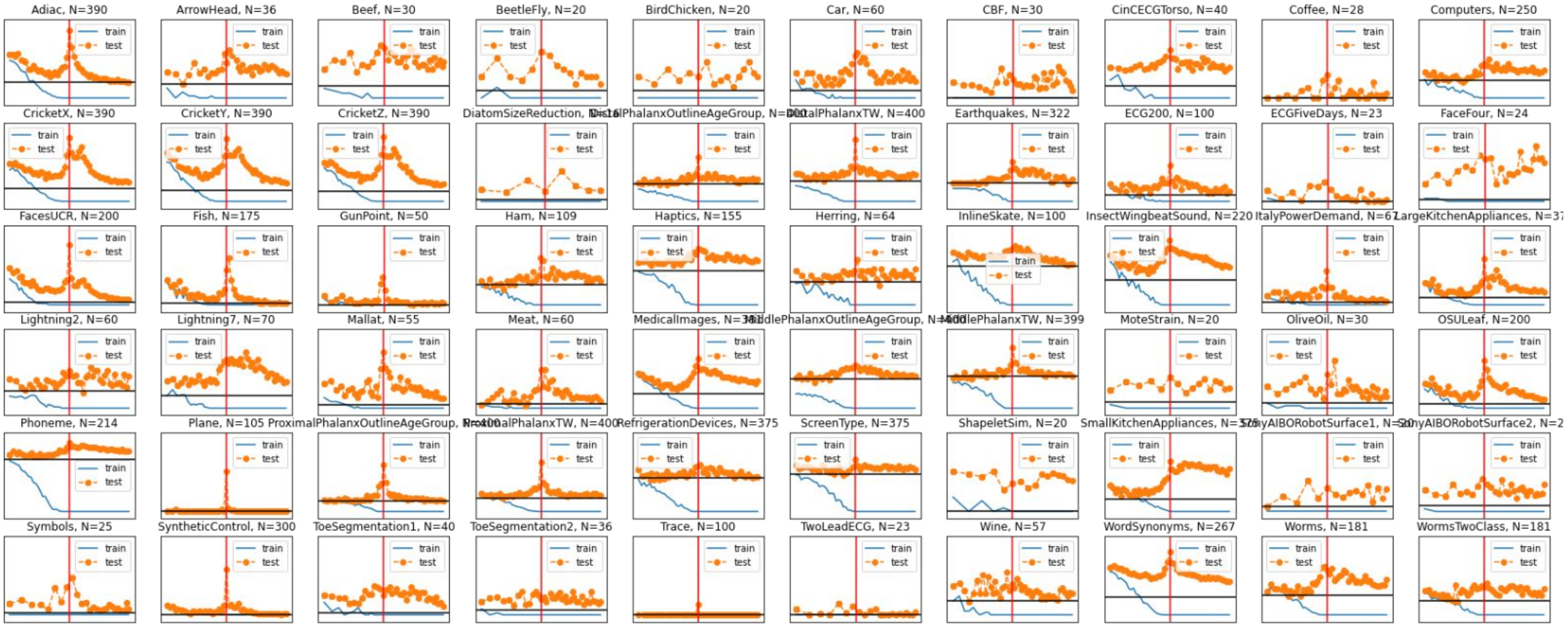


Random Relu features



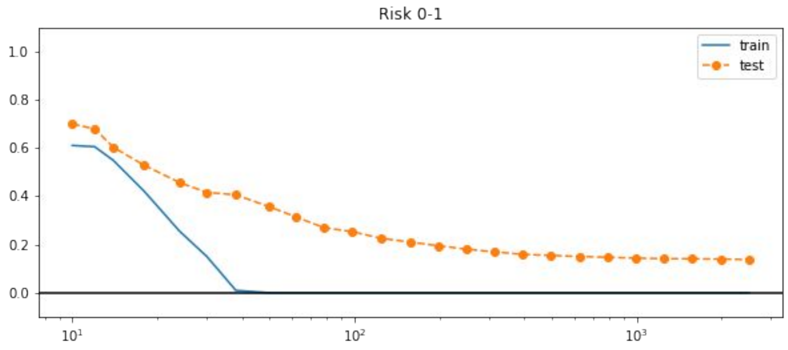
Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.

Rocket

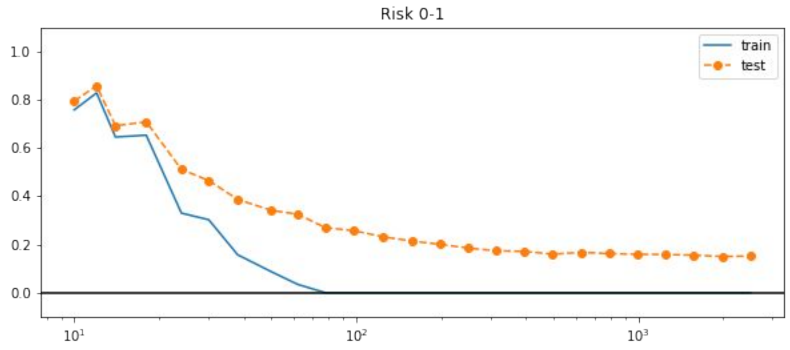

Adiac, N=390
ArrowHead, N=36
Beef, N=30
BeetleFly, N=20
BirdChicken, N=20
Car, N=60
CBF, N=30
CinCECGTorso, N=40
Coffee, N=28
Computers, N=250
CricketX, N=390
CricketY, N=390
CricketZ, N=390
Earthquakes, N=322
ECG200, N=100
ECGFiveDays, N=23
FaceFour, N=24
FacesUCR, N=200
Fish, N=175
GunPoint, N=50
Ham, N=109
Haptics, N=155
Herring, N=64
InlineSkate, N=100
InsectWingbeatSound, N=220
Lightning2, N=60
Lightning7, N=70
Mallat, N=55
Meat, N=60
MoteStrain, N=20
OliveOil, N=30
OSULeaf, N=200
Phoneme, N=214
Plane, N=105
RefrigerationDevices, N=375
ScreenType, N=375
ShapeletSim, N=20
Symbols, N=25
SyntheticControl, N=300
ToeSegmentation1, N=40
ToeSegmentation2, N=36
Trace, N=100
TwoLeadECG, N=23
Wine, N=57
WordSynonyms, N=267
Worms, N=181
WormsTwoClass, N=181
train
test

Different models

Logistic Regression



Linear SVM

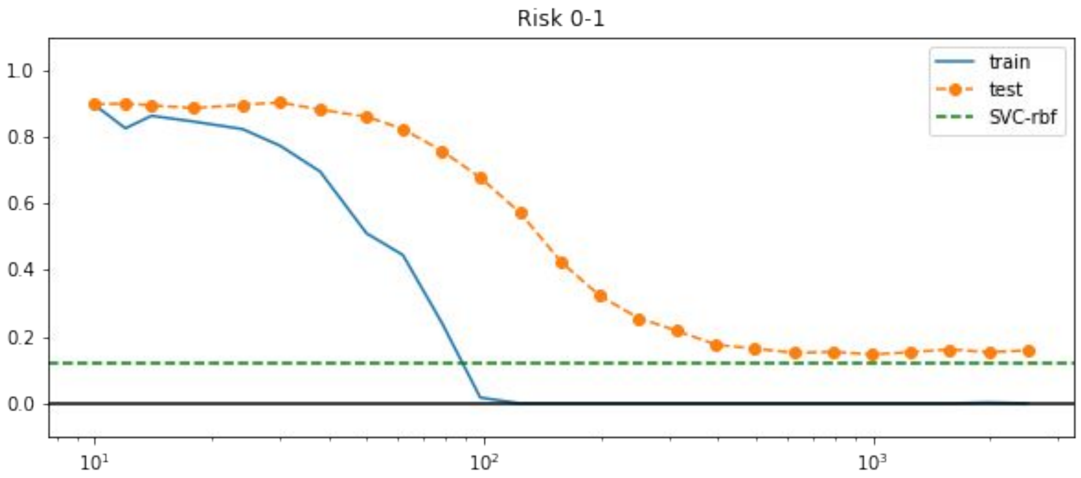


Linear SVM with RFF

$$\boldsymbol{\omega} \sim p(\boldsymbol{\omega})$$
$$b \sim \text{Uniform}(0, 2\pi)$$
$$z_{\boldsymbol{\omega}}(\mathbf{x}) = \sqrt{2}\cos(\boldsymbol{\omega}^\top \mathbf{x} + b).$$

$$\mathbf{z}(\mathbf{x}) = \begin{bmatrix} \frac{1}{\sqrt{R}} z_{\boldsymbol{\omega}_1}(\mathbf{x}) \\ \frac{1}{\sqrt{R}} z_{\boldsymbol{\omega}_2}(\mathbf{x}) \\ \vdots \\ \frac{1}{\sqrt{R}} z_{\boldsymbol{\omega}_R}(\mathbf{x}) \end{bmatrix}$$

$$\begin{aligned}\mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{y}) &= \frac{1}{R}\sum_{r=1}^{R} z_{\boldsymbol{\omega}_r}(\mathbf{x}) z_{\boldsymbol{\omega}_r}(\mathbf{y}) \\ &= \frac{1}{R}\sum_{r=1}^{R} 2\cos(\boldsymbol{\omega}_r^\top \mathbf{x} + b_r)\cos(\boldsymbol{\omega}_r^\top \mathbf{y} + b_r) \\ &= \frac{1}{R}\sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^\top (\mathbf{x} - \mathbf{y})) \\ &\approx \mathbb{E}_{\boldsymbol{\omega}}[\cos(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{y}))] \\ &= k(\mathbf{x}, \mathbf{y}).\end{aligned}$$



Rahimi, Ali, and Benjamin Recht. "Random features for large-scale kernel machines." *Advances in neural information processing systems* 20 (2007).