Generalization Performance of the Random Fourier Features Method

Anna C. Gilbert

University of Michigan

joint work with Yitong Sun (Univ. of Michigan)

SVM Preliminaries: two label classification

- Data with labels: Let $\mathcal{X} \subset \mathbb{R}^d$ and data $(x, y) \in \mathcal{X} \times \{-1, 1\}$, generated by the distribution $\mathcal{D}$.

- Classification task: Find/approximate the Bayes classifier

$$f(x) = \operatorname{sgn}\left(\mathop{\mathbb{P}}_{(X,Y)\sim\mathcal{D}} \{Y = 1 \mid X = x\} - \frac{1}{2} \right).$$

- Use kernelized SVM:

$$\operatorname{sgn}(\langle w, \phi(x) \rangle_{\mathcal{H}} + b),$$

where $\phi : \mathcal{X} \to \mathcal{H}$ satisfies

$$\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y).$$

SVM Preliminaries: solution of SVM

- Primal:

$$\hat{w} = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \frac{C}{m} \sum_{j=1}^{m} \max\left(0, 1 - y_j \left(\langle w, \phi\left(x_j\right)\rangle_{\mathcal{H}} + b\right)\right) + \frac{1}{2} \|w\|_{\mathcal{H}}^2 \ . \tag{1}$$

- Dual:

$$\hat{\alpha} = \underset{\substack{\forall j\, 0 \le \alpha_j \le 1/m \\ \sum_j \alpha_j y_j = 0}}{\operatorname{argmax}} C \sum_{j=1}^{m} \alpha_j - \frac{C^2}{2} \left(\alpha \circ y\right)^{\mathsf{T}} K_m \left(\alpha \circ y\right) \ ,$$

  where $\alpha \circ y = \left(\alpha_1 y_1, \ldots, \alpha_m y_m\right)^{\mathsf{T}}$ and $K_m = \left[k\left(x_i, x_j\right)\right]$.

- Then,

$$\hat{w} = C \sum_{j=1}^{m} \hat{\alpha}_j y_j \phi\left(x_j\right) \ .$$

SVM Preliminaries: equivalent formulation

- Primal:

$$\hat{w} = \underset{\|w\|_{\mathcal{H}} \leq \Lambda}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} \max\left(0, 1 - y_j \left(\langle w, \phi(x_j) \rangle_{\mathcal{H}} + b\right)\right). \quad (2)$$

- Dual:

$$\hat{\alpha} = \underset{\substack{0 \leq \alpha_j \leq 1/m \\ \sum_j \alpha_j y_j = 0}}{\operatorname{argmax}} \sum_{j=1}^{m} \alpha_j - \Lambda \sqrt{(\alpha \circ y)^{\mathsf{T}} K_m (\alpha \circ y)} \quad (3)$$

- When

$$\Lambda = C \sqrt{(\hat{\alpha} \circ y)^{\mathsf{T}} K_m (\hat{\alpha} \circ y)},$$

  the solution is the same with (1).
- Force $b = 0$ for technical simplicity and then corresponding dual problem is of the same form except for dropping the constraint $\sum_j \alpha_j y_j = 0$.

RFF: Radial Basis Function Kernels

- Radial Basis Function kernels (RBF):

$$k(x, y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}.$$

- Have the following feature map $\phi : \mathcal{X} \to L_2\left(\mathbb{R}^d, g, \gamma\right)$

$$\phi(x) = e^{ig^\intercal x}.$$

  where γ is the standard norm distribution on $\mathbb{R}^d$.

- Easy to verify that

$$\mathop{\mathbb{E}}_{g\sim\gamma}\left[e^{\frac{ig^\intercal x}{\sigma}}\overline{e^{\frac{ig^\intercal y}{\sigma}}}\right] = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}.$$

RFF: Approximate features

- To speed up computation, reduce the dimension of $(\operatorname{span}\{\phi(x_j)\})$ to $N \ll m$.
- [Rahimi and Recht, '08] proposed to approximate ϕ by $\tilde{\phi} : \mathcal{X} \to \mathbb{R}^{2N}$ with

$$\tilde{\phi}(x) = \frac{1}{\sqrt{N}} \left(\cos\left(\frac{g_1^\intercal x}{\sigma}\right), \sin\left(\frac{g_1^\intercal x}{\sigma}\right), \ldots, \cos\left(\frac{g_N^\intercal x}{\sigma}\right), \sin\left(\frac{g_N^\intercal x}{\sigma}\right) \right)^\intercal$$

  where g_k's are i.i.d. from $N(0, I_d)$.

- Then as $N \to \infty$,

$$\tilde{k}(x, y) := \tilde{\phi}(x)^\intercal \tilde{\phi}(y) \longrightarrow \mathop{\mathbb{E}}_{g \sim \gamma} [e^{ig^\intercal x} \overline{e^{ig^\intercal y}}] = k(x, y)$$

SVM Using Random Features

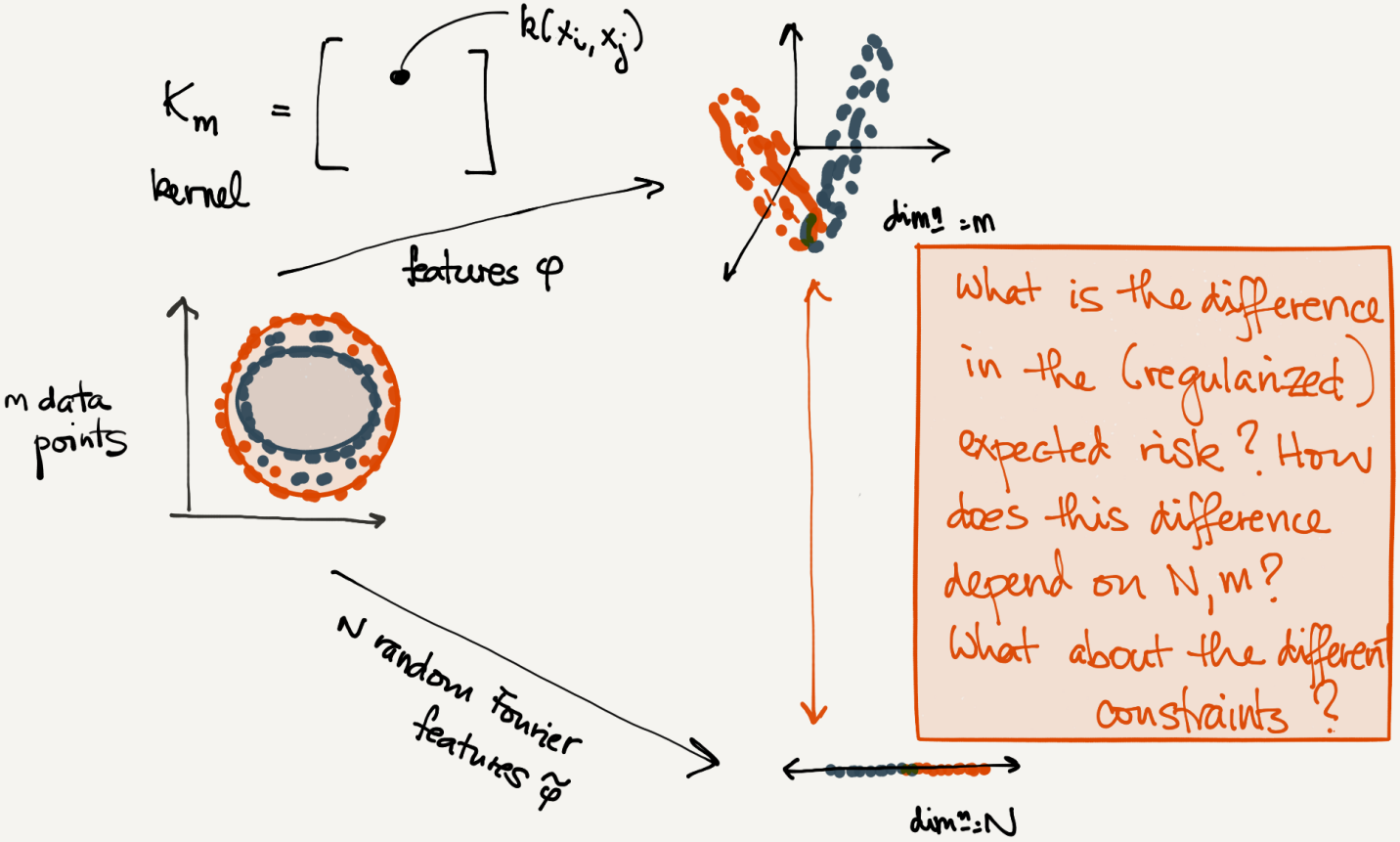
- Denote

$$\tilde{w} = \underset{\|w\|_2 \leq \tilde{\Lambda}}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} \max\left(0, 1 - y_j w^{\mathsf{T}} \tilde{\phi}(x_i)\right) \tag{4}$$

$$\tilde{\alpha} = \underset{0 \leq \alpha_j \leq 1/m}{\operatorname{argmax}} \sum_{j=1}^{m} \alpha_j - \tilde{\Lambda} \sqrt{(\alpha \circ y)^{\mathsf{T}} \tilde{K}_m (\alpha \circ y)}, \tag{5}$$

where $\tilde{K}_m = \left[\tilde{k}(x_i, x_j)\right]$ and $\tilde{C} = \tilde{\Lambda} \sqrt{(\tilde{\alpha} \circ y)^{\mathsf{T}} \tilde{K}_m (\tilde{\alpha} \circ y)}$

- Question: how large should N be so that we do not lose too much accuracy? In what quantity?


K_m =
$k(x_i, x_j)$
kernel
features φ
$\dim^{\varphi} = m$
m data points
N random fourier features $\tilde{\varphi}$
$\dim^{m} = N$
What is the difference in the (regularized) expected risk? How does this difference depend on N, m? What about the different constraints?

Related work: How close are the sep'ors? $\|\hat{w} - \tilde{w}\|$

- Alas, $\hat{w}$ and $\tilde{w}$ are in different spaces!
- [Cortes, et al., 2010] mapped them to the same ambient space and showed

$$\|\hat{w} - \tilde{w}\|_2^2 \leq \frac{C}{\sqrt{m}} \|K_m\|^{1/2} \left\|\tilde{K}_m - K_m\right\|^{1/2}, ,$$

  where C is the common regularization parameter.
- By a matrix Bernstein inequality [Tropp, 2015],

$$\left\|\tilde{K}_m - K_m\right\| \leq \left(\frac{4m\|K_m\|}{N} \log \frac{2m}{\delta}\right)^{1/2}.$$

- Unfortunately, $\|K_m\| = O(m)$ in most cases.

Related work: What about the true risk? $R(\tilde{w}) - R(\hat{w})$

- A better indicator of the performance of a model is the true (expected) risk

$$R(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(y, h(x)).$$

  where, for SVM classification, ℓ is usually hinge loss.
- Hypothesis class for kSVM is $h(x) = \langle w, \phi(x)\rangle_{\mathcal{H}} + b$
- Regularized (expected) risk

$$\mathbf{R}(\langle w, \phi(x)\rangle_{\mathcal{H}}) = R(\langle w, \phi(x)\rangle_{\mathcal{H}}) + \frac{1}{2C}\|w\|_{\mathcal{H}}^2$$

Related work: What about the true risk? $R(\tilde{w}) - R(\hat{w})$

- [Rahimi and Recht, 2009] showed that

$$R(\tilde{w}) - R(\hat{w}) \leq O\left(\frac{C}{\sqrt{m}} + \frac{C}{\sqrt{N}}\right) \quad \text{with high probability.}$$

- Result is, however, for a special regularlization setup (instead of 2-norm regularization): $\tilde{w}$ and $\hat{w}$ are optimal solutions on

$$\left\{ C \sum_j \alpha_j \tilde{\phi}(x_j) \,\middle|\, \|\alpha\|_\infty \leq \frac{1}{m} \right\} \text{ and } \left\{ C \sum_j \alpha_j \phi(x_j) \,\middle|\, \|\alpha\|_\infty \leq \frac{1}{m} \right\}$$

- Does not match the 2-norm regularlized ERM formulation!

Related work: others

- [Bach, 2015] considers constrained ERM problems that involve Lipschitz loss functions in a more general setting. The result on the error rate requires the random features be generated from the optimized density $q(v)$, which minimizes

$$d_{\max}(q, \lambda) = \sup_{v \in \mathcal{V}} \frac{1}{q(v)} \langle \varphi(v, \cdot), (\Sigma + \lambda I)^{-1} \varphi(v, \cdot) \rangle_{L_2(d\rho)} .$$

  However, the optimized $q(v)$ depends on the distribution of data, which is in general not available. Our work corresponds to $q(v) = 1$. Would need to compute $d_{\max}(1, \lambda)$.
- [Rudi, 2016] considers ridge regression under RFF, not classification.

Our work: cast of characters

	Accurate model	**Approximate model**
Feature	$\phi(x) \in \mathcal{H}$	$\tilde{\phi}(x) \in \tilde{\mathcal{H}} = \mathbb{R}^N$
Gram matrix	k, K_m	$\tilde{k}, \tilde{K}_m$
Regularization parameters	C	$\tilde{C}$
Norm constraint	Λ	$\tilde{\Lambda}$
Primal solution	$\hat{w}$	$\tilde{w}$
Dual solution	$\hat{\alpha}$	$\tilde{\alpha}$

Table: Summary of notations for accurate models and approximate models

Our work: summary

Reference	Excess risk	Performance indicator	Regularization parameters
Theorem 1	$O\left(\frac{1}{\sqrt{N}}\right)$	regularized expected risk	$\tilde{C} = C$
Corollary 2	$O\left(\left(\frac{\|K_m\|}{Nm}\right)^{1/2} + \left(\frac{\|K_m\|^7}{Nm^5}\right)^{1/8}\right)$	expected risk	$\tilde{C} = C$
Theorem 3	$O\left(\frac{1}{N^{1/4}}\right)$	expected risk	$\tilde{\Lambda} = \Lambda$
Theorem 4	$O\left(\frac{1}{\sqrt{N}}\right)$	expected risk	$\tilde{\Lambda} = \eta_1\Lambda$ or $\tilde{C} = \eta_2 C$

Summary of generalization performance of random features method under different regularization contraints

Our work: Lemma (regularized empirical risk)

Lemma

Assume $|\phi(x;\omega)| \leq \kappa$, $\|\tilde{\phi}(x)\|_2 \leq \tilde{\kappa}$, *and* $\tilde{C} = C$. *Let* $0 < \delta < 1$. *Then, with probability at least* $1 - \delta$,

$$\mathbf{R}_m^C\left(\langle \tilde{w}, \tilde{\phi}(x)\rangle_{\tilde{\mathcal{H}}}\right) - \mathbf{R}_m^C\left(\langle \hat{w}, \phi(x)\rangle_{\mathcal{H}}\right) \leq C\left(\frac{\kappa^2 \|K_m\|}{Nm} \log\frac{2m}{\delta}\right)^{1/2}.$$

Proof.

Each term on the left hand side is the minimum of the regularized empirical risk. Use strong duality and subtract dual forms of LHS

$$\begin{aligned}
\mathbf{R}_m(\tilde{w}) - \mathbf{R}_m(\hat{w}) &\leq \sup_{0\leq\alpha_i\leq\frac{1}{m}} \frac{C}{2}\left|(\alpha\circ\mathbf{y})^{\mathsf{T}}\left(K_m - \tilde{K}_m\right)(\alpha\circ\mathbf{y})\right| \\
&\leq \frac{C}{2m}\left\|K_m - \tilde{K}_m\right\| \\
&\leq C\left(\frac{\kappa^2\|K_m\|}{Nm}\log\frac{2m}{\delta}\right)^{1/2},
\end{aligned}$$

with probability at least $1 - \delta$. The last inequality comes from the result of the matrix Bernstein inequality. □

Our work: Theorem 1 (regularized expected risk)

Theorem (1)

Same assumptions as Lemma. Then, with probability at least $1-\delta$,

$$\mathbf{R}_{\mathcal{D}}^{C}\left(\langle \tilde{w}, \tilde{\phi}(x)\rangle_{\tilde{\mathcal{H}}}\right) - \mathbf{R}_{\mathcal{D}}^{C}\left(\langle \hat{w}, \phi(x)\rangle_{\mathcal{H}}\right) \leq C_1 \left(\frac{1}{m}\right)^{1/2} + C_2 \left(\frac{\|K_m\|}{Nm}\right)^{1/2},$$

where C_1 *and* C_2 *are polynomials of* κ, $\tilde{\kappa}$, L, C, $\log m$ *and* $\log \delta$. *Explicit constants are shown in the proof.*

Proof.

Use Lemma + classical trick in estimating excess risk (expected - empirical). $\square$

Our work: risk gap for norm constraints: $\Lambda = \tilde{\Lambda}$

Theorem (3)

$\hat{w}$ and $\tilde{w}$ are solutions to exact and approximate models (Eqns (2) and (3)). Assume that $\Lambda = \tilde{\Lambda}$. Then

$$R(\tilde{w}) - R(\hat{w}) \leq O\left(\frac{\Lambda}{N^{1/4}}\right), \text{ with high probability.}$$

Our work: Choose parameters carefully

Theorem (4)

Assume $\hat{w}$ and $\tilde{w}$ are solutions to exact and approximate models (Eqns (2) and (3)) and

$$\frac{\tilde{\Lambda}}{\Lambda} \geq \eta = \left(\frac{(\hat{\alpha} \circ y)^{\mathsf{T}} \tilde{K}_m (\hat{\alpha} \circ y)}{(\hat{\alpha} \circ y)^{\mathsf{T}} K_m (\hat{\alpha} \circ y)} \right)^{1/2},$$

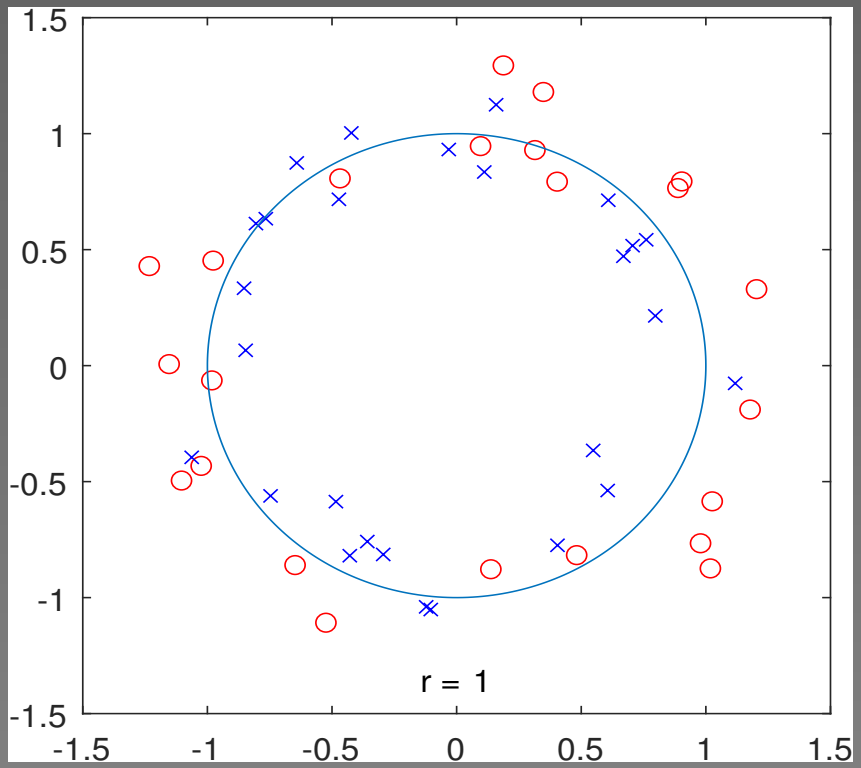
where $\hat{\alpha}$ is the dual solution. Then

$$R(\tilde{w}) - R(\hat{w}) \leq O\left(\frac{C}{\sqrt{N}} + \frac{\eta \Lambda}{\sqrt{m}} \right) \leq O\left(\frac{C}{\sqrt{N}} + \frac{C}{\sqrt{m}} \right),$$

with high probability.

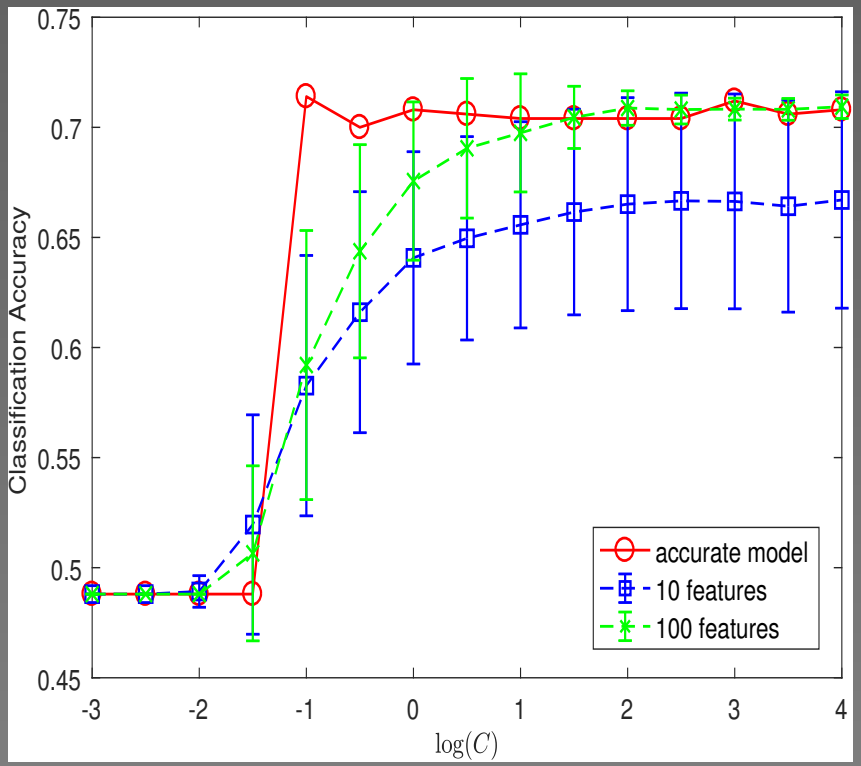
Experiments: regularization versus random features

Figure: Distribution of training samples. 50 out of 1000 points are shown in the graph. Blue crosses represent the points generated by the uniform distribution over the annulus $0.65 \leq R \leq 1.15$. Red circles represent the points generated by the uniform distribution over the annulus $0.85 \leq R \leq 1.35$. The unit circle is the best classifier for these two distributions. Obviously, the best classifier can achieve at most 70% classification accuracy due to the overlap between two distributions.



Experiments: regularization versus random features

Figure: Performance of the LSVM with random Fourier features. Each curve plots the classification accuracy on the test dataset of the hypothesis learned by the model with different choice of C. Larger values of $\log(C)$ indicate weaker regularization. The error bars on the approximate model with random features represent the mean and standard deviation of 50 runs.



Further Questions

- In practice, the choice of regularization parameter C is always greater than the sample size m. In such a case, $O(C/\sqrt{N})$ is meaningless.
- Theorem (4) holds only when the random features model allows smaller margin controlled by η. What is the scale of η?
- When $\tilde{\Lambda}/\Lambda = \eta$, $\tilde{C}/C$ is not 1. This may explain why [Cortes, 2010] failed to obtain an informative upper bound for $\|\tilde{w} - \hat{w}\|$.
- Based on the results of [Cortes, 2010] and Theorem (3), it seems additional requirements like $C = \tilde{C}$ or $\Lambda = \tilde{\Lambda}$ only enlarge the gap between approximate and accurate models. Then, what are the general criteria that we should follow when comparing two different learning models?