Learning Kernels with Random Features

Aman Sinha & John Duchi, Stanford University
{amans, jduchi}@stanford.edu

Introduction

- Random features efficiently approximate kernel machines, but require a user-defined kernel choice
- We extend the random feature approach to kernel learning
- Our workflow is simple: create random features with an arbitrary kernel, solve a fast optimization problem to select a subset, then train a model with the selected features
- Compared to other methods, we get competitive performance at a fraction of the cost

Problem Setup and Approach

Learning a kernel

Input: n datapoints $(x^i, y^i) \in \mathbb{R}^d \times \{-1, 1\}$
Problem: Consider a kernel $K_Q(x, x') = \mathbb{E}_Q[\phi(x; W)\phi(x'; W)]$. We want a good Q

- Try kernel alignment: $\text{maximize}_{Q \in \mathcal{P}} \ \sum_{i,j} K_Q(x^i, x^j) y^i y^j$
- Constrain Q with power f-divergences around a base distribution:

$$f(t) = t^k - 1 \ (k \geq 2), \quad \mathcal{P} := \{Q : D_f(Q \| P_0) \leq \rho\}$$

- Approximate the problem with random features:

$$\underset{q \in \mathcal{P}_{N_w}}{\text{maximize}} \ \sum_{i,j} y^i y^j \sum_{m=1}^{N_w} q_m \phi(x^i, w^m)\phi(x^j, w^m) \tag{1}$$

where $w^i \overset{\text{iid}}{\sim} P_0$ and $\mathcal{P}_{N_w} := \{q : D_f(q \| \mathbf{1}/N_w) \leq \rho\}$

- Fast (near-linear time) solution, often results in sparse q

Using the learned kernel

Empirical risk minimization in 2 possible flavors:

- Standard random features with new distribution q
  - Draw D new samples $w^i \overset{\text{iid}}{\sim} q$
  - Define features $\phi^i := [\phi(x^i, w^1) \ \cdots \ \phi(x^i, w^D)]^T$ and solve

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \left\{ \sum_{i=1}^n c\left(\tfrac{1}{\sqrt{D}} \theta^T \phi^i, y^i\right) + r(\theta) \right\} \tag{2}$$

- Nonrandom features with new distribution q
  - Use original $w^i \overset{\text{iid}}{\sim} P_0$ from optimization above
  - Define features $\phi^i := [\phi(x^i, w^1) \ \cdots \ \phi(x^i, w^{N_w})]^T$ and solve

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \left\{ \sum_{i=1}^n c(\theta^T \text{diag}(\widehat{q})^{\frac{1}{2}} \phi^i, y^i) + r(\theta) \right\} \tag{3}$$

  - Efficient for sparse q, useful when desired # of features $D \geq \text{nnz}(q)$

Consistency & Generalization Performance

Consistency

- The solution to problem (1) approaches a population optimum as data and random sampling increase ($n \to \infty$ and $N_w \to \infty$)
- Consider the slightly more general setting: let $S : \mathcal{X} \times \mathcal{X} \to [-1, 1]$, $S_{ij} := S(x^i, x^j)$ (with $y \in \{-1, 1\}$, we have $S_{ij} = y^i y^j$)
- Define alignment functions

$$T(P) := \mathbb{E}[S(X, X') K_P(X, X')], \qquad \widehat{T}(P) := \tfrac{1}{n(n-1)} \textstyle\sum_{i \neq j} S_{ij} K_P(x^i, x^j)$$

Theorem 1. *Let $\widehat{Q}_w$ maximize $\widehat{T}(Q)$ over $Q \in \mathcal{P}_{N_w}$. Let $C_\rho = \frac{2(\rho+1)}{\sqrt{1+\rho}-1}$ and $D_\rho = \sqrt{8(1+\rho)}$. Then, with probability at least $1 - 3\delta$ over the sampling of both (x, y) and W, we have*

$$\left| T(\widehat{Q}_w) - \sup_{Q \in \mathcal{P}} T(Q) \right| \leq 4C_\rho \sqrt{\frac{\log(2N_w)}{N_w}} + D_\rho \sqrt{\frac{\log \frac{2}{\delta}}{N_w}} + 2D_\rho \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

Generalization performance

- The estimator (3) uses the function class

$$\mathcal{F}_{N_w} := \left\{ h(x) = \textstyle\sum_{m=1}^{N_w} \alpha_m \sqrt{q_m} \phi(x, w^m) \mid q \in \mathcal{P}_{N_w}, \|\alpha\|_2 \leq B \right\}$$

- Define the true misclassification risk and ν-empirical misclassification risk for an estimator h as

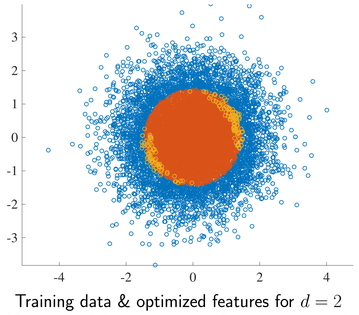
$$R(h) := \mathbb{P}(Yh(X) < 0), \qquad \widehat{R}_\nu(h) := \tfrac{1}{n} \textstyle\sum_{i=1}^n \min\left\{1, \left[1 - yh(x^i)/\nu\right]_+\right\}$$

Theorem 2. *The generalization performance for an estimator from $\mathcal{F}_{N_w}$ satsfies $\sup_{h \in \mathcal{F}_{N_w}} \{R(h) - \widehat{R}_\nu(h)\} \leq \frac{2}{\nu} B \sqrt{\frac{2(1+\rho)}{n}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$ with probability at least $1 - \delta$.*

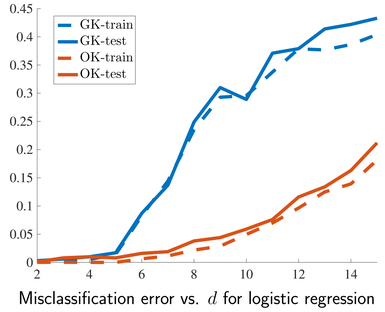
Empirical Evaluations

Learning a new kernel with a poor choice of P_0

- Synthetic data $x^i \overset{\text{iid}}{\sim} \mathsf{N}(0, I) \in \mathbb{R}^d$, $y_i = \text{sign}(\|x\|_2 - \sqrt{d})$
- Use P_0 corresponding to Gaussian kernel (ill-suited for this data)



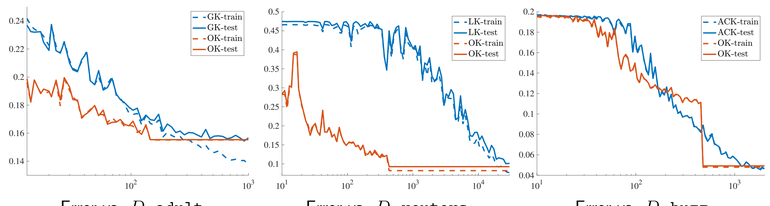
Training data & optimized features for $d = 2$
(Data in red/blue, optimized features in yellow)



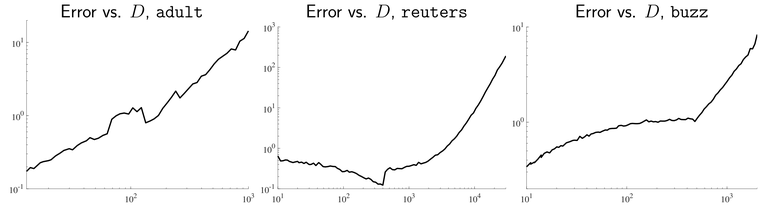
Misclassification error vs. d for logistic regression
(GK=Gaussian kernel, OK=optimized kernel)

Benchmark datasets

- Compare with standard random features on `adult`, `reuters`, `buzz`
- Employ (2) when $D < \text{nnz}(q)$, (3) otherwise
- Use Gaussian, linear, arc-cosine kernel respectively



Error vs. D, `adult` | Error vs. D, `reuters` | Error vs. D, `buzz`



Speedup vs. D, `adult` | Speedup vs. D, `reuters` | Speedup vs. D, `buzz`

Best test results

Dataset	n	d	Model	Our error (%), time(s)		Random error (%), time(s)	
`adult`	32561	123	Logistic	15.54	3.6	15.44	43.1
`reuters`	23149	47236	Ridge	9.27	0.8	9.36	295.9
`buzz`	105530	77	Ridge	4.92	2.0	4.58	11.9

- Our method is faster at moderate to large D and shows better performance than standard random features at small to moderate D

Comparison with joint optimization

- Consider joint optimization via SVM dual

$$\begin{array}{ll} \text{minimize}_{q \in \mathcal{P}_{N_w}} \sup_\alpha & \alpha^T \mathbf{1} - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y^i y^j \sum_{m=1}^{N_w} q_m \phi(x^i, w^m)\phi(x^j, w^m) \\ & \text{subject to } \mathbf{0} \preceq \alpha \preceq C\mathbf{1}, \ \ \alpha^T y = 0 \end{array}$$

- Use subsampled data due to computational cost

Performance on subsampled data

Dataset	Our error (%), time(s)		Joint error (%), time(s)	
`adult`	16.36	1.8	16.31	198.1
`reuters`	9.66	0.6	8.96	173.3
`buzz`	8.32	0.4	7.08	137.5

- Our method's efficiency outweighs marginal loss in accuracy

Conclusion

- We exploit computational advantages of random features to develop a fast, scalable kernel-learning optimization procedure
- Concentration bounds guarantee our procedure is consistent, and our estimator generalizes well
- Emprirical results indicate we learn new structure, and we attain competitive results faster than other methods

A. Sinha is supported by a Hertz Fellowship and a Stanford Graduate Fellowship.