Random Fourier features

Random Fourier features (Rahim and Recht, 2007) scale shift-invariant kernels to large numbers of inputs by using linear models on \( z(x) \), where \( z : \mathbb{R}^d \rightarrow \mathbb{R}^d \) has \( k(x,y) \approx z(x)^T z(y) \).

Let \( \Delta := x - y, \) and \( k(x,y) = k(\Delta), k(0) \) be a continuous PSD kernel. Its Fourier transform \( P(\omega) \) is a probability measure (Bochner’s theorem).

One embedding is:

\[
\phi(x) := \frac{1}{\sqrt{D}} \begin{pmatrix}
\sin(\omega_1^T x) \\
\cos(\omega_1^T x) \\
\vdots \\
\sin(\omega_D^T x) \\
\cos(\omega_D^T x)
\end{pmatrix}, \quad \omega_i \overset{\text{iid}}{\sim} P(\omega).
\]

\( \phi(x)^T \psi(y) \) is an average of \( D/2 \) terms \( \cos(\omega_i^T \Delta) \); note \( \mathbb{E} \cos(\omega_i^T \Delta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \Delta^2} dP(\omega) = k(\Delta) \).

Another has more samples from \( P(\omega) \), but with additional non-shift-invariant noise:

\[
\psi(x) := \frac{1}{\sqrt{D}} \begin{pmatrix}
\cos(\omega_1^T x + b_1) \\
\sin(\omega_1^T x + b_1) \\
\vdots \\
\cos(\omega_D^T x + b_D) \\
\sin(\omega_D^T x + b_D)
\end{pmatrix}, \quad \omega_i \overset{\text{iid}}{\sim} P(\omega), \quad b_i \overset{\text{iid}}{\sim} \text{Unif}[-\pi,\pi],
\]

\( \psi(x)^T \psi(y) \) is the mean of \( D \) terms of the form \( \cos(\omega_i^T \Delta) + \cos(\omega_i^T (x+y) + 2b_i) \).

\[ \phi \] is better for Gaussian kernels

For \( k(\Delta) := \exp(-||\Delta||^2 / (2\sigma^2)) \),

\[
\begin{aligned}
\text{Var} & \cos(\omega^T \Delta) = \frac{1}{2} \left( 1 - \exp\left(-\frac{||\Delta||^2}{2\sigma^2}\right) \right) \leq \frac{1}{2}.
\end{aligned}
\]

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image1.png}
\caption{The variance per dimension for the Gaussian kernel. The difference in variance is higher for larger kernel values.}
\end{figure}

Improved uniform convergence

We can tighten the bound for \( \phi \) and show one for \( \psi \).

• Let \( \ell \) be the diameter of the domain \( \mathcal{X} \subseteq \mathbb{R}^d \).

• Let \( \sigma_p^2 := \mathbb{E}||\omega||^2, \sigma_w^2 := \sup_\Delta \left[ 2 \text{Var} \cos(\omega^T \Delta) \right] \).

Define \( f_{\phi}(x,y) := \phi(x)^T \phi(y) - k(x,y) \) to be the error for \( \phi \), and let \( \alpha_\epsilon := \min\left( 1, \frac{\sigma_p^2}{\ell^2 \sigma_w^2} + \frac{\epsilon}{4\epsilon} \right) \); then

\[
\Pr\left( \|f_\phi\|_\infty \geq \epsilon \right) \leq \beta_\delta \left( \frac{\sigma_p^2 \ell^2}{\epsilon} \right)^{1/4} \exp\left( -\frac{D \epsilon^2}{8(d+2)\alpha_\epsilon} \right).
\]

For \( \psi \), define \( f_{\psi}(x,y) := \psi(x)^T \psi(y) - k(x,y) \) as well as \( \alpha'_\epsilon := \min\left( 1, \frac{\sigma_p^2}{\ell^2 (1 + \sigma_w^2)} + \frac{\epsilon}{4\epsilon} \right) \); then

\[
\Pr\left( \|f_\psi\|_\infty \geq \epsilon \right) \leq \beta'_\delta \left( \frac{\sigma_p^2 \ell^2}{\epsilon} \right)^{1/4} \exp\left( -\frac{D \epsilon^2}{32(d+2)\alpha'_\epsilon} \right).
\]

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image2.png}
\caption{The coefficients \( \beta_\delta \) (blue, for \( \phi \)) and \( \beta'_\delta \) (orange, for \( \psi \)). The bound for \( \phi \) is always tighter than that for \( \psi \).
\end{figure}

\section{Numerical results with \( d = 1 \)}

Gaussian kernel, \( \sigma = 1 \). \( \phi \) has solid lines, \( \psi \) dashed.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image3.png}
\caption{Average max error within a given radius.}
\end{figure}

\section{Further results}

The paper also has:

• Exact expectations and concentration bounds of squared \( L_2 \) error, for any measure.

• Bounds on changes in the outputs of ridge regression, SVM, and maximum mean discrepancy tests due to the features.

• More experiments.

This work was funded in part by DARPA grant FA87501280324. D.I.S. is also supported by a Sandia Campus Executive Program fellowship.