Weighted Sums of Random Kitchen Sinks

based on Rahimi & Recht, NIPS 2008

Dino Sejdinovic

Gatsby Unit, UCL

May 9, 2014

Notation

- Fit a function $f : \mathcal{X} \to \mathbb{R}$ to examples $\{(x_i, y_i)\}_{i=1}^{m} \overset{i.i.d.}{\sim} P$ with $y_i \in \{-1, +1\}$.

Notation

- Fit a function $f : \mathcal{X} \to \mathbb{R}$ to examples $\{(x_i, y_i)\}_{i=1}^m \overset{i.i.d.}{\sim} P$ with $y_i \in \{-1, +1\}$.
- f is assumed to be of the form $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$, where
  - $\phi(x;\omega)$ are the feature functions / nonlinearities / weak learners, parametrized by $\omega \in \Omega$ ($\tanh(\omega^\top x)$, $\cos\left(\omega^\top x + b\right)$)
  - $\alpha : \Omega \to \mathbb{R}$ are the weights associated to each of the individual feature functions

Notation

- Fit a function $f : \mathcal{X} \to \mathbb{R}$ to examples $\{(x_i, y_i)\}_{i=1}^m \overset{i.i.d.}{\sim} P$ with $y_i \in \{-1, +1\}$.
- f is assumed to be of the form $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$, where
  - $\phi(x;\omega)$ are the feature functions / nonlinearities / weak learners, parametrized by $\omega \in \Omega$ ($\tanh(\omega^\top x)$, $\cos\left(\omega^\top x + b\right)$)
  - $\alpha : \Omega \to \mathbb{R}$ are the weights associated to each of the individual feature functions
- Loss function $c(f(x), y)$ determines the *empirical risk*:

$$\mathbf{R}_{\mathrm{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} c(f(x_i), y_i)$$

Notation

- Fit a function $f : \mathcal{X} \to \mathbb{R}$ to examples $\{(x_i, y_i)\}_{i=1}^m \overset{i.i.d.}{\sim} P$ with $y_i \in \{-1, +1\}$.
- f is assumed to be of the form $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$, where
  - $\phi(x;\omega)$ are the feature functions / nonlinearities / weak learners, parametrized by $\omega \in \Omega$ ($\tanh(\omega^\top x)$, $\cos\left(\omega^\top x + b\right)$)
  - $\alpha : \Omega \to \mathbb{R}$ are the weights associated to each of the individual feature functions
- Loss function $c(f(x), y)$ determines the *empirical risk*:

$$\mathbf{R}_{\text{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} c(f(x_i), y_i)$$

- *True risk*:

$$\mathbf{R}[f] = \mathbb{E}_P c(f(X), Y)$$

- Typically: fix the number of non-linearities K and minimize the empirical risk over both the parameters ω and the weights α:

$$\min_{\omega_1,\ldots,\omega_K;\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}}\left[\sum_{k=1}^{K} \alpha_k \phi(\cdot;\omega_k)\right].$$

- Typically: fix the number of non-linearities K and minimize the empirical risk over both the parameters ω and the weights α:

$$\min_{\omega_1,\ldots,\omega_K;\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}} \left[\sum_{k=1}^{K} \alpha_k \phi(\cdot;\omega_k) \right].$$

- **This paper**: pick non-linearities randomly and optimize only over the weights:

$$\{\omega_k\}_{k=1}^{K} \overset{i.i.d.}{\sim} \pi; \qquad \min_{\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}} \left[\sum_{k=1}^{K} \alpha_k \phi(\cdot;\omega_k) \right].$$

- Typically: fix the number of non-linearities K and minimize the empirical risk over both the parameters ω and the weights α:

$$\min_{\omega_1,\ldots,\omega_K;\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}}\left[\sum_{k=1}^{K}\alpha_k\phi(\cdot;\omega_k)\right].$$

- **This paper**: pick non-linearities randomly and optimize only over the weights:

$$\{\omega_k\}_{k=1}^{K} \overset{i.i.d.}{\sim} \pi; \qquad \min_{\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}}\left[\sum_{k=1}^{K}\alpha_k\phi(\cdot;\omega_k)\right].$$

- Linear least squares or linear SVMs on the featurized inputs:

$$x_i \quad \mapsto \quad z(x_i) := [\phi(x_i;\omega_1)\cdots\phi(x_i;\omega_K)]^\top.$$

- Typically: fix the number of non-linearities K and minimize the empirical risk over both the parameters ω and the weights α:

$$\min_{\omega_1,\ldots,\omega_K;\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}}\left[\sum_{k=1}^{K} \alpha_k \phi(\cdot;\omega_k)\right].$$

- **This paper**: pick non-linearities randomly and optimize only over the weights:

$$\{\omega_k\}_{k=1}^{K} \overset{i.i.d.}{\sim} \pi; \qquad \min_{\alpha_1,\ldots,\alpha_K} \frac{1}{m}\sum_{i=1}^{m} c\left(\sum_{k=1}^{K} \alpha_k \phi(x_i;\omega_k), y_i\right).$$

- Linear least squares or linear SVMs on the featurized inputs:

$$x_i \quad \mapsto \quad z(x_i) := \left[\phi(x_i;\omega_1)\cdots\phi(x_i;\omega_K)\right]^\top.$$

- Typically: fix the number of non-linearities K and minimize the empirical risk over both the parameters ω and the weights α:

$$\min_{\omega_1,\ldots,\omega_K;\alpha_1,\ldots,\alpha_K} \mathsf{R}_{\mathrm{emp}}\left[\sum_{k=1}^{K}\alpha_k\phi(\cdot;\omega_k)\right].$$

- **This paper**: pick non-linearities randomly and optimize only over the weights:

$$\{\omega_k\}_{k=1}^{K} \overset{i.i.d.}{\sim} \pi; \qquad \min_{\alpha_1,\ldots,\alpha_K} \frac{1}{m}\sum_{i=1}^{m} c\left(\alpha^\top z(x_i), y_i\right).$$

- Linear least squares or linear SVMs on the featurized inputs:

$$x_i \quad \mapsto \quad z(x_i) := \left[\phi(x_i;\omega_1)\cdots\phi(x_i;\omega_K)\right]^\top.$$

- Typically: fix the number of non-linearities K and minimize the empirical risk over both the parameters ω and the weights α:

$$\min_{\omega_1,\ldots,\omega_K;\alpha_1,\ldots,\alpha_K} \mathbf{R}_{\text{emp}}\left[\sum_{k=1}^{K}\alpha_k\phi(\cdot;\omega_k)\right].$$

- **This paper**: pick non-linearities randomly and optimize only over the weights:

$$\{\omega_k\}_{k=1}^{K} \overset{i.i.d.}{\sim} \pi; \qquad \min_{\alpha_1,\ldots,\alpha_K} \frac{1}{m}\sum_{i=1}^{m} c\left(\alpha^\top z(x_i), y_i\right) \quad \text{s.t. } \|\alpha\|_\infty \leq \frac{C}{K}.$$

- Linear least squares or linear SVMs on the featurized inputs:

$$x_i \quad \mapsto \quad z(x_i) := \left[\phi(x_i;\omega_1)\cdots\phi(x_i;\omega_K)\right]^\top.$$

Greedy function approximations

Given function f^* and a probability measure μ (to measure fidelity):

$$\begin{aligned}
(\omega_k, \alpha_k) &= \arg\min_{\omega_k, \alpha_k} \left\| (1-\alpha_k)\hat{f}_{k-1} + \alpha_k \phi(\cdot;\omega_k) - f^* \right\|_{L^2(\mu)} \\
\hat{f}_k &\leftarrow (1-\alpha_k)\hat{f}_{k-1} + \alpha_k \phi(\cdot;\omega_k)
\end{aligned}$$

Uniform bounds for functions in a given smoothness class. For example (Jones, 1992; Barron, 1993) if $f^* = \sum_{k=1}^{\infty} \alpha_k^* \phi(\cdot;\omega_k^*)$,

$$\left\| \hat{f}_K - f^* \right\|_{L^2(\mu)} = O\left(\frac{\|\alpha\|_1}{\sqrt{K}} \right)$$

Space $\mathcal{F}_\pi$

- Functions of interest $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$ are endowed with the norm w.r.t. sampling distribution π:

$$\|f\|_\pi = \sup_{\omega\in\Omega} \frac{|\alpha(\omega)|}{\pi(\omega)}$$

- Space of interest $\mathcal{F}_\pi = \left\{f = \int \alpha(\omega)\phi(\cdot;\omega)d\omega \mid \|f\|_\pi < \infty\right\}$.
- $|\alpha(\omega)| \leq C\pi(\omega)$: weights α decay more rapidly than π.
- Smoothness class induced by the *sampling distribution* π.

The richness of space $\mathcal{F}_\pi$

- Define kernel $k(x, y) = \mathbb{E}_{\omega \sim \pi} [\phi(x; \omega)\phi(y; \omega)]$

The richness of space $\mathcal{F}_\pi$

- Define kernel $k(x,y) = \mathbb{E}_{\omega\sim\pi}\left[\phi(x;\omega)\phi(y;\omega)\right]$
- Then RKHS $\mathcal{H}_k$ consists of functions $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$ such that $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)}d\omega < \infty$

The richness of space $\mathcal{F}_\pi$

- Define kernel $k(x,y) = \mathbb{E}_{\omega\sim\pi}\left[\phi(x;\omega)\phi(y;\omega)\right]$
- Then RKHS $\mathcal{H}_k$ consists of functions $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$ such that $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)}d\omega < \infty$
- $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)^2}\pi(\omega)d\omega \leq \|f\|^2_\pi$, so $\mathcal{F}_\pi \subseteq \mathcal{H}_k$

The richness of space $\mathcal{F}_\pi$

- Define kernel $k(x, y) = \mathbb{E}_{\omega \sim \pi}[\phi(x; \omega)\phi(y; \omega)]$
- Then RKHS $\mathcal{H}_k$ consists of functions $f(x) = \int \alpha(\omega)\phi(x; \omega)d\omega$ such that $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)} d\omega < \infty$
- $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)^2}\pi(\omega)d\omega \leq \|f\|^2_\pi$, so $\mathcal{F}_\pi \subseteq \mathcal{H}_k$
- The case $\phi(x; \omega) = \cos\left(w^\top x + b\right)$ covers all translation-invariant kernels: $\pi(w)$ is then the inverse Fourier transform of $\kappa(x) = k(x, 0)$.

The richness of space $\mathcal{F}_\pi$

- Define kernel $k(x,y) = \mathbb{E}_{\omega\sim\pi}\left[\phi(x;\omega)\phi(y;\omega)\right]$
- Then RKHS $\mathcal{H}_k$ consists of functions $f(x) = \int \alpha(\omega)\phi(x;\omega)d\omega$ such that $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)}d\omega < \infty$
- $\|f\|^2_{\mathcal{H}_k} = \int \frac{\alpha(\omega)^2}{\pi(\omega)^2}\pi(\omega)d\omega \le \|f\|^2_\pi$, so $\mathcal{F}_\pi \subseteq \mathcal{H}_k$
- $\mathcal{F}_\pi$ is **dense** in $\mathcal{H}_k$: it contains all functions of the form

$$
\begin{aligned}
f(x) = \sum_{i=1}^m a_i k(x_i, x) &= \sum_{i=1}^m a_i \int \phi(x_i;\omega)\phi(x;\omega)\pi(\omega)d\omega \\
&= \int \left[\underbrace{\pi(\omega)\sum_{i=1}^m a_i\phi(x_i;\omega)}_{\alpha(\omega)}\right]\phi(x;\omega)d\omega,
\end{aligned}
$$

since $\frac{|\alpha(\omega)|}{\pi(\omega)} \le \sum_{i=1}^m |a_i| < \infty$.

Hypothesis space

- After randomization of $\{\omega_k\}_{k=1}^K$ we find the best function in a random subspace spanned by $\{\phi(x;\omega_k)\}_{k=1}^K$

$$\hat{\mathcal{F}}_\omega = \left\{ f = \sum_{k=1}^{K} \alpha_k \phi(\cdot;\omega_k) \,\middle|\, |\alpha_k| \leq \frac{C}{K} \right\}$$

- Two sources of error:
  - Approximation: is the risk of the best function in $\hat{\mathcal{F}}_\omega$ close to the risk of the best function in C-ball of $\mathcal{F}_\pi$?
  - Estimation: is the empirical risk in $\hat{\mathcal{F}}_\omega$ close to the true risk?

Given function f^* and a probability measure μ (to measure fidelity):

- sample $\{\omega_k\}_{k=1}^K \overset{i.i.d.}{\sim} \pi$, batch-fit α's:

$$\alpha = \arg\min_{\alpha} \left\| \sum_{k=1}^K \alpha_k \phi(\cdot;\omega_k) - f^* \right\|_{L^2(\mu)}$$

- (**Lemma 1**): Now, w.p. $1-\delta$

$$\left\| \hat{f}_K - f^* \right\|_{L^2(\mu)} = O\left(\frac{\|f^*\|_\pi}{\sqrt{K}} \left(1 + \sqrt{2\log\frac{1}{\delta}} \right) \right)$$

- So uniform result only over the balls in $\mathcal{F}_\pi$

Main result

Theorem

Suppose that $\sup_{x,\omega} |\phi(x;\omega)| \le 1$ *and that* $c(f(x), y) = c\,(f(x)y)$ *depends only on the product* $f(x)y$ *and is L-Lipschitz. Let* π *be any distribution on* Ω*. Then random featurization with* $\{\omega_k\}_{k=1}^{K} \overset{i.i.d.}{\sim} \pi$ *gives w.p.* $1 - 2\delta$*:*

$$\mathbf{R}[\hat{f}] - \min_{\|f\|_\pi \le C} \mathbf{R}[f] \quad \le \quad O\left(LC \left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{K}} \right) \sqrt{\log \frac{1}{\delta}} \right).$$

Approximation error: Lemma 1

Lemma

Let $f^* \in \mathcal{F}_\pi$ with $\|f^*\|_\pi \leq C$, and $\{\omega_k\}_{k=1}^K \overset{i.i.d.}{\sim} \pi$. Then there exists $\hat{f} = \sum_{k=1}^K \hat{\alpha}_k \phi(\cdot;\omega_k)$, with $|\hat{\alpha}_k| \leq \frac{C}{K}$, s.t. w.p. $1-\delta$:

$$\left\|\hat{f}_K - f^*\right\|_{L^2(\mu)} \leq \frac{C}{\sqrt{K}}\left(1+\sqrt{2\log\frac{1}{\delta}}\right).$$

Proof.

Denote $f^* = \int \alpha^*(\omega)\phi(\cdot;\omega)d\omega$ and let $f_k = \frac{\alpha^*(\omega_k)}{\pi(\omega_k)}\phi(\cdot;\omega_k)$. Now $\mathbb{E}_{\omega_k} f_k = \int \frac{\alpha^*(\omega_k)}{\pi(\omega_k)}\phi(\cdot;\omega_k)\pi(\omega_k)d\omega_k = f^*$. Define $\hat{f}_K = \frac{1}{K}\sum_{k=1}^K f_k$, i.e., weights are $\hat{\alpha}_k = \frac{\alpha^*(\omega_k)}{K\pi(\omega_k)}$, and clearly $|\hat{\alpha}_k| \leq \frac{C}{K}$. Moreover, $\|f_k\|_{L^2(\mu)} \leq C$ a.s. and the proof follows by the concentration around the mean of the empirical average $\frac{1}{K}\sum_{k=1}^K f_k$ in $L^2(\mu)$. □

Approximation error

Lemma

$$\mathbf{R}[\hat{f}_K] - \mathbf{R}[f^*] \leq \frac{LC}{\sqrt{K}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right).$$

$$\begin{aligned}
\mathbf{R}[\hat{f}_K] - \mathbf{R}[f^*] &= \mathbb{E}_P\left[c\left(\hat{f}_K(x)y\right) - c\left(f^*(x)y\right)\right] \\
&\leq \mathbb{E}_P\left|c\left(\hat{f}_K(x)y\right) - c\left(f^*(x)y\right)\right| \\
\text{(cis Lipschitz)} \quad &\leq L\mathbb{E}_P\left|\left(\hat{f}_K(x) - f^*(x)\right)y\right| \\
(|y| \leq 1) \quad &\leq L\mathbb{E}_{P_X}\left|\hat{f}_K(x) - f^*(x)\right| \\
\text{(Jensen)} \quad &\leq L\sqrt{\mathbb{E}_{P_X}\left(\hat{f}_K(x) - f^*(x)\right)^2} = L\left\|\hat{f}_K - f^*\right\|_{L^2(P_X)}
\end{aligned}$$

Summary

- Selecting many random non-linearities can achieve better accuracy-time tradeoff than greedy algorithms that optimize both non-linearities and their weights
  - Much more non-linearities required
  - Optimization much much faster
- Assuming that the sampling distribution has thicker tails than the weight of the target, approximation error decays as $O(1/\sqrt{K})$ with K randomly sampled non-linearities
- Constant depends on the "smoothness" of target w.r.t. sampling distribution, so the result is not uniform on target space.