Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contents

1

Monte Carlo Methods

1.1 Monte Carlo methods

Let $\lambda \geq 0$ be a probability density function on $G \subset \mathbb{R}^d$ such that

$$\int_G \lambda(\omega) d\omega = 1. \tag{1.1}$$

For example:

$$\lambda(\omega) = \frac{1}{|G|}, \tag{1.2}$$

if G is bounded. The expectation is defined:

$$\mathbb{E} g := \int_G g(\omega)\lambda(\omega) d\omega \tag{1.3}$$

and for any $h = h(\omega_1, \omega_2, \ldots, \omega_n) : G \times G \cdots G \mapsto \mathbb{R}$

$$\mathbb{E}_n h := \int_{G\times G\times \ldots \times G} h(\omega_1, \omega_2, \ldots, \omega_n)\lambda(\omega_1)\lambda(\omega_2)\ldots\lambda(\omega_n) d\omega_1 d\omega_2 \ldots d\omega_n. \tag{1.4}$$

1.1.1 A basic result

Lemma 1. *For any $g \in L^\infty(G)$, we have*

$$\mathbb{E}_n \Big(\mathbb{E} g - \frac{1}{n}\sum_{i=1}^n g(\omega_i)\Big)^2 = \begin{cases} \frac{1}{n}\mathbb{E}((\mathbb{E} g - g)^2) \leq \frac{1}{n}\sup_{\omega,\omega\in G} |g(\omega) - g(\omega')|^2 \\ \frac{1}{n}\Big(\mathbb{E}(g^2) - (\mathbb{E}(g))^2\Big) \leq \frac{1}{n}\mathbb{E}(g^2) \leq \frac{1}{n}\|g\|^2_{L^\infty}, \end{cases} \tag{1.5}$$

Proof. First note that

$$
\begin{aligned}
\left(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i)\right)^2 &= \frac{1}{n^2}\left(n\mathbb{E}g - \sum_{i=1}^{n} g(\omega_i)\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}(\mathbb{E}g - g(\omega_i))\right)^2 \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}(\mathbb{E}g - g(\omega_i))(\mathbb{E}g - g(\omega_j)) \\
&= \frac{I_1}{n^2} + \frac{I_2}{n^2}.
\end{aligned} \tag{1.6}
$$

with

$$
I_1 = \sum_{i=1}^{n}(\mathbb{E}g - g(\omega_i))^2, \quad I_2 = \sum_{i\neq j}^{n}((\mathbb{E}g)^2 - \mathbb{E}(g)(g(\omega_i) + g(\omega_j)) + g(\omega_i)g(\omega_j))). \tag{1.7}
$$

Consider I_1, for any i,

$$
\mathbb{E}_n(\mathbb{E}g - g(\omega_i))^2 = \mathbb{E}_n(\mathbb{E}g - g)^2.
$$

Thus,

$$
\mathbb{E}_n I_1 = n\mathbb{E}((\mathbb{E}g - g)^2).
$$

For I_2, note that

$$
\mathbb{E}_n g(\omega_i) = \mathbb{E}_n g(\omega_j) = \mathbb{E}(g)
$$

and, for $i \neq j$,

$$
\begin{aligned}
\mathbb{E}_n(g(\omega_i)g(\omega_j)) &= \int_{G\times G\times\ldots\times G} g(\omega_j)g(\omega_j)\lambda(\omega_1)\lambda(\omega_2)\ldots\lambda(\omega_n)d\omega_1 d\omega_2\ldots d\omega_n \\
&= \int_{G\times G} g(\omega_j)g(\omega_j)\lambda(\omega_1)\lambda(\omega_1)\lambda(\omega_2)d\omega_1 d\omega_2 \\
&= \mathbb{E}_n(g(\omega_i))\mathbb{E}_n(g(\omega_j)) = [\mathbb{E}(g)]^2.
\end{aligned} \tag{1.8}
$$

Thus

$$
\mathbb{E}_n(I_2) = \mathbb{E}_n\left(\sum_{i\neq j}^{n}((\mathbb{E}g)^2 - \mathbb{E}(g)(\mathbb{E}(g(\omega_i)) + \mathbb{E}(g(\omega_j))) + \mathbb{E}(g(\omega_i)g(\omega_j)))\right) = 0. \tag{1.9}
$$

Consequently, there exist the following two formulas for $\mathbb{E}_n\left(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i)\right)^2$:

$$
\mathbb{E}_n\left(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i)\right)^2 = \frac{1}{n^2}\mathbb{E}_n I_1 = \begin{cases} \frac{1}{n}\mathbb{E}((\mathbb{E}g - g)^2) \\ \frac{1}{n}(\mathbb{E}(g^2) - (\mathbb{E}g)^2). \end{cases} \tag{1.10}
$$

Based on the first formula above, since

$$
|g(\omega) - \mathbb{E}g| = |\int_G (g(\omega) - g(\tilde{\omega}))\lambda(\tilde{\omega})d\tilde{\omega}| \le \sup_{\omega,\omega\in G}|g(\omega) - g(\omega')|,
$$

it holds that

$$\mathbb{E}_n\left(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i)\right)^2 \le \frac{1}{n}\sup_{\omega,\omega\in G} |g(\omega) - g(\omega')|^2. \tag{1.11}$$

Due to the second formula above,

$$\mathbb{E}_n\left(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i)\right)^2 \le \frac{1}{n}\mathbb{E}(g^2) \le \frac{1}{n}\|g\|_{L^\infty}^2 \tag{1.12}$$

which completes the proof. □

Of course, we can use this to prove a high probability result.

Corollary 1. *Under the assumptions of the preceding lemma, we have*

$$\bar{\mathbb{P}}[(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i))^2 > \frac{k}{n}\|g\|_{L^\infty}^2] < \frac{1}{k} \tag{1.13}$$

Proof.

$$\bar{\mathbb{P}}[(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i))^2 > \epsilon] \le \epsilon^{-1}\bar{\mathbb{E}}(\mathbb{E}g - \frac{1}{n}\sum_{i=1}^{n} g(\omega_i))^2 \le \frac{1}{n\epsilon}\|g\|_{L^\infty}^2. \tag{1.14}$$

□

This corollary implies that the set of ω_i where the estimate $n^{-1}\sum_{i=1}^{n} g(\omega_i)$ is far from the desired value $\mathbb{E}g$ is small.

The practical usefulness of this algorithm depends upon the existence of a *repeatable* process (for instance some physical process) which *generates ω according to a desired distribution μ.*

The precise meaning of this last statement is essentially that the strong law of large numbers holds. Specifically, if $\omega_1, ..., \omega_n, ...$ is a infinite sequence generated by the process, and $A \subset \Omega$ is any a measurable set, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \chi_A(\omega_i) = \mu(A). \tag{1.15}$$

Generating n independent samples means generating $\omega_1, ..., \omega_n$ from μ^n according to the above notion. The existence of a realizable process generating samples from a probability distribution, and the practical use of such processes is an interesting topic in the intersection of statistics, physics, and computer science. In addition, statistics/probability theory studies how to take samples from one probability distribution and transform them to samples from another distribution.

1.1.2 Application

Lemma 2. *Let*

$$f(x) = \int_G g(x,\theta)\lambda(\theta)d\theta = \mathbb{E}(g) \tag{1.16}$$

with $\lambda(\theta) \geq 0$ *and* $\|\lambda(\theta)\|_{L^1(G)} = 1$. *For any* $n \geq 1$, *there exist* $\theta_i^* \in G$ *such that*

$$\|f - f_n\|^2_{L^2(\Omega)} \leq \frac{1}{n}\int_G \|g(\cdot,\theta)\|^2_{L^2(\Omega)}\lambda(\theta)d\theta = \frac{1}{n}\mathbb{E}(\|g(\cdot,\theta)\|^2_{L^2(\Omega)})$$

where $\|g(\cdot,\theta)\|^2_{L^2(\Omega)} = \int_\Omega [g(x,\theta)]^2 d\mu(x)$, *and*

$$f_n(x) = \frac{1}{n}\sum_{i=1}^n g(x,\theta_i^*).$$

Proof. Introducing a probability distribution $\lambda(\theta)$,

$$f(x) = \mathbb{E}(g). \tag{1.17}$$

By Lemma 1,

$$\mathbb{E}_n\left(\left(\mathbb{E}(g(x,\cdot)) - \frac{1}{n}\sum_{i=1}^n g(x,\theta_i)\right)^2\right) \leq \frac{1}{n}\mathbb{E}(g^2)$$

and

$$\mathbb{E}_n\left(h(\theta_1,\theta_2,\cdots,\theta_n)\right) \leq \frac{1}{n}\mathbb{E}\Big(\int_\Omega g^2 d\mu(x)\Big),$$

by taking integral where

$$h(\theta_1,\theta_2,\cdots,\theta_n) = \int_\Omega \left(\mathbb{E}(g(x,\cdot)) - \frac{1}{n}\sum_{i=1}^n g(x,\theta_i)\right)^2 d\mu(x).$$

Sine $\mathbb{E}_n(1) = 1$ and $\mathbb{E}_n(h) \leq \frac{1}{n}\mathbb{E}\Big(\int_\Omega g^2 d\mu(x)\Big)$, there exists $(\theta_1^*,\theta_2^*,\cdots,\theta_n^*) \in G\times G\times\cdots\times G$ such that

$$h(\theta_1^*,\theta_2^*,\cdots,\theta_n^*) \leq \frac{1}{n}\int_\Omega \mathbb{E}(g^2)d\mu(x).$$

Otherwise, $\mathbb{E}_n(h) > \frac{1}{n}\mathbb{E}\Big(\int_\Omega g^2 d\mu(x)\Big)$ if $h(\theta_1,\theta_2,\cdots,\theta_n) > \frac{1}{n}\int_\Omega \mathbb{E}(g^2))d\mu(x)$. This implies that

$$\|f - f_n\|^2_{L^2(\Omega)} \leq \frac{1}{n}\int_G \|g(\cdot,\theta)\|^2_{L^2(\Omega)}\lambda(\theta)d\theta,$$

which completes the proof. □

We also have a more general version of the above lemma.

Lemma 3. *Let*

$$f(x) = \int_G g(x,\theta)\lambda(\theta)d\theta = \mathbb{E}(g) \tag{1.18}$$

with $\|\lambda(\theta)\|_{L^1(\Theta)} = 1$. *For any* $n \geq 1$, *there exist* $\theta_i^* \in G$ *such that*

$$\|f - f_n\|^2_{H^m(\Omega)} \leq \int_G \|g(\cdot,\theta)\|^2_{H^m(\Omega)}\lambda(\theta)d\theta = \frac{1}{n}\mathbb{E}(\|g(\cdot,\theta)\|^2_{H^m(\Omega)})$$

where

$$f_n(x) = \frac{1}{n}\sum_{i=1}^n g(x,\theta_i^*)$$

In particular, if

$$|D^\alpha g(x,\theta)| \leq C, \quad \forall x,\theta,|\alpha| \leq m \tag{1.19}$$

Then

$$\|f - f_n\|_{H^m(\Omega)} \leq \binom{m+d}{m}^{1/2} |\Omega|^{1/2} n^{-1/2}.$$

For any $f(x) = \int_G g(x,\theta)\rho(\theta)d\theta$ with $\|\rho\|_{L^1(\Theta)} \neq 1$. Let $\lambda(\theta) = \frac{\rho(\theta)}{\|\rho\|_{L^1(\Theta)}}$. Thus,

$$f(x) = \|\rho\|_{L^1(\Theta)} \int_G g(x,\theta)\lambda(\theta)d\theta \tag{1.20}$$

with $\|\lambda(\theta)\|_{L^1(\Theta)} = 1$. We can apply the above two lemmas to the given function $f(x)$.

1.2 Integral representations of functions

1.2.1 Fourier representation

Consider the Fourier transform:

$$\hat{f}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega\cdot x} f(x) dx \quad \forall \omega \in \mathbb{R}^d. \tag{1.21}$$

We write $\hat{f}(\omega) = e^{i\theta(\omega)}|\hat{f}(\omega)|$. By Fourier inversion formula,

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega\cdot x}\hat{f}(\omega)d\omega = \int_{\mathbb{R}^d} e^{i(\omega\cdot x+\beta(\omega))}|\hat{f}(\omega)|d\omega. \tag{1.22}$$

Since $f(x)$ is real-valued, it implies that, for x

$$\begin{aligned} f(x) &= \mathrm{Re}\int_{\mathbb{R}^d} e^{i\omega\cdot x}\hat{f}(\omega)d\omega \\ &= \mathrm{Re}\int_{\mathbb{R}^d} e^{i\omega\cdot x}e^{i\beta(\omega)}|\hat{f}(\omega)|d\omega \\ &= \int_{\mathbb{R}^d} \cos(\omega\cdot x+\beta(\omega))|\hat{f}(\omega)|d\omega. \end{aligned} \tag{1.23}$$

Then we have

$$f(x) = \int_{\mathbb{R}^d} k(x,\omega)d\omega, \tag{1.24}$$

with

$$k(x,\omega) = \cos(\omega\cdot x+\beta(\omega))|\hat{f}(\omega)| \tag{1.25}$$

and

$$|k(x,\omega)| \le |\hat{f}(\omega)| = \rho(\omega). \tag{1.26}$$

Theorem 1. *There exist $\omega_i \in \mathbb{R}^d$, s.t., $G = \mathbb{R}$ and*

$$\int_\Omega (f(x) - f_n(x))^2 \le \frac{1}{n}\int_{\mathbb{R}^d} |\hat{f}(\omega)|d\omega, \tag{1.27}$$

where

$$f_n(x) = \frac{\|\hat{f}\|_{L^1}}{n}\sum_{i=1}^n \frac{cos(\omega_i^*\cdot x+\beta_i^*)}{\rho(\omega_i^*)}. \tag{1.28}$$

Note that

$$f_n = \sum_{i=1}^n \frac{cos(\omega_i^*\cdot x+\beta_i^*)}{\rho(\omega_i^*)} \in {}_n\mathrm{N}(\sigma,n), \tag{1.29}$$

with

$$\sigma(t) = \cos(t). \tag{1.30}$$

1.2.2 Double Fourier representation

Assume that σ is a locally Riemann integrable function and $\sigma \in L^1(\mathbb{R})$ and thus the Fourier transform of σ is well-defined and continuous. Since σ is non-zero and

$$\hat{\sigma}(\omega) = \frac{1}{2\pi}\int_{\mathbb{R}} \sigma(t)e^{-i\omega t}dt, \tag{1.31}$$

this implies that $\hat{\sigma}(a) \neq 0$ for some $a \neq 0$. Via a change of variables $t = w \cdot x + b$ and $dt = db$, this means that for all x and ω, we have

$$\begin{aligned} 0 \neq \hat{\sigma}(a) &= \frac{1}{2\pi}\int_{\mathbb{R}} \sigma(\omega \cdot x + b)e^{-ia(\omega\cdot x+b)}db \\ &= e^{-ia\omega\cdot x}\frac{1}{2\pi}\int_{\mathbb{R}} \sigma(\omega \cdot x + b)e^{-ia+b}db, \end{aligned} \tag{1.32}$$

and so

$$e^{ia\omega\cdot x} = \frac{1}{2\pi\hat{\sigma}(a)}\int_{\mathbb{R}} \sigma(\omega \cdot x + b)e^{-iab}db. \tag{1.33}$$

Likewise, since the growth condition also implies that $\sigma^{(k)} \in L^1$, we can differentiate the above expression under the integral with respect to x.

This allows us to write the Fourier mode $e^{ia\omega\cdot x}$ as an integral of neuron output functions. We substitute this into the Fourier representation of f (note that the assumption we make implies that $\hat{f} \in L^1$ so this is rigorously justified for a.e. x) to get

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega\cdot x}\hat{f}(\omega)d\omega = \int_{\mathbb{R}^d}\int_{\mathbb{R}} \frac{1}{2\pi\hat{\sigma}(a)}\sigma\left(a^{-1}\omega \cdot x + b\right)\hat{f}(\omega)e^{-iab}dbd\omega = \int_{\mathbb{R}^d\times\mathbb{R}} k(x,\theta)d\theta \tag{1.34}$$

where $\theta = (\omega, b)$ and

$$k(x,\theta) = \frac{1}{2\pi\hat{\sigma}(a)}\sigma\left(a^{-1}\omega \cdot x + b\right)\hat{f}(\omega)e^{-iab}.$$

Thus we have

$$\begin{aligned} |k(x,\theta)| &\leq \frac{1}{2\pi|\hat{\sigma}(a)|}\max_{x\in\Omega}|\sigma\left(a^{-1}\omega \cdot x + b\right)||\hat{f}(\omega)| \\ &\leq h(\omega, b)|\hat{f}(\omega)| = \rho(\theta) \end{aligned}, \tag{1.35}$$

where

$$h(\omega, b) = \max_{x\in\Omega}|\sigma(a^{-1}\omega \cdot x + b)|. \tag{1.36}$$

if we ignore the coefficient. Thus, following the discussion in last section, the next step is to analyze $h(\omega, b)$ which we will discuss in the next section. Before that, let us introduce a special case of the above representation once the activation function is periodic.

References

[1] L. Chen, P. Sun, and J. Xu. Optimal anisotropic meshes for minimizing interpolation errors in l^p-norm. *Mathematics of Computation*, 76(257):179–204, 2007.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[4] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[5] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

[6] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, 54(190):483–493, 1990.

[7] J. Xu. Finite element methods. Lecture notes, 2020.