Scaling Laws of Optimization Notes

KC Tsiolis

November 14, 2024

The purpose of Bach's article[1] is to illustrate the scaling law of the risk for unconstrained convex quadratic optimization problems with respect to number of gradient descent iterations k. Early on in the article, Bach points out that scaling laws in statistics have been present for a long time. If we ignore the optimization algorithm and consider only the number of features d and the dataset size n, then we have the canonical $\frac{\sigma^2 d}{n}$ bound for ordinary least squares (OLS).

The optimization problem is set up as follows. Suppose we have $F : \mathbb{R}^d \to \mathbb{R}$ defined by

$$F(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_*) + F_*, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{d\times d}$ is positive semi-definite (PSD). Hence, $\boldsymbol{\theta}_*$ is a minimizer of F (unique if $\mathbf{H} \succ 0$) and F_* is the minimum value. Then, a GD step takes the form

$$\begin{aligned} \boldsymbol{\theta}_k &= \boldsymbol{\theta}_{k-1} - \gamma \nabla F(\boldsymbol{\theta}_{k-1}) \\ &= \boldsymbol{\theta}_{k-1} - \gamma \mathbf{H}(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_*) \end{aligned} \tag{2}$$

As a consequence,

$$\begin{aligned} \boldsymbol{\theta}_k - \boldsymbol{\theta}_* &= (\mathbf{I} - \gamma \mathbf{H})(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_*) \\ &= (\mathbf{I} - \gamma \mathbf{H})^k(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*). \end{aligned} \tag{3}$$

Hence, the optimality gap (for function values) is

$$F(\boldsymbol{\theta}_k) - F_* = \frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T \mathbf{H}(\mathbf{I} - \gamma \mathbf{H})^{2k}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*), \tag{4}$$

since $\mathbf{H}$ and $(\mathbf{I} - \gamma \mathbf{H})$ commute.

"Strongly convex analysis". Bach points out that the classical "strongly convex analysis" employs crude bounds by assuming that $\lambda_{\min}(\mathbf{H}) = \mu > 0$ and $\lambda_{\max}(\mathbf{H}) = L$. If we take $\gamma = 1/L$,

[1] `https://francisbach.com/scaling-laws-of-optimization/`

all eigenvalues of $\mathbf{I} - \gamma\mathbf{H}$ are nonnegative and the largest is $1 - \frac{\mu}{L}$. Then,

$$\begin{aligned} F(\boldsymbol{\theta}_k) - F_* &\leq \frac{1}{2}\left(1 - \frac{\mu}{L}\right)^{2k} (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T \mathbf{H} (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*) \\ &= \left(1 - \frac{\mu}{L}\right)^{2k} \big(F(\boldsymbol{\theta}_0) - F_*\big). \end{aligned} \tag{5}$$

Hence, in the strongly convex case, we have exponential convergence (though the constant approaches 1 as $\mu \to 0$).

"Non-strongly convex analysis". Alternatively, we have a "non-strongly convex" analysis noting that $\mathbf{H}(\mathbf{I} - \gamma\mathbf{H})^{2k}$ has eigenvalues of the form $\lambda_i(1 - \gamma\lambda_i)^{2k}$ (where λ_i are eigenvalues of $\mathbf{H}$), we can use the fact that the function $\alpha \mapsto \alpha(1-\alpha)^{2k}$ is maximized at $\alpha = \frac{1}{2k+1}$. Hence,

$$\frac{1}{\gamma}\gamma\lambda_i(1 - \gamma\lambda_i)^{2k} \leq \frac{1}{\gamma}\frac{1}{2k+1}\left(\frac{2k}{2k+1}\right)^{2k} \leq \frac{1}{2\gamma e k}. \tag{6}$$

Hence, if we take $\gamma = 1/L$,

$$F(\boldsymbol{\theta}_k) - F_* \leq \frac{L}{4e}||\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*||_2^2 \cdot \frac{1}{k}. \tag{7}$$

Consider the first two plots in Bach's blog post to appreciate how "bad" these bounds are.

Laplace's method applied to power-law covariance. For simplicity, assume that $\mathbf{H} = \mathrm{diag}(h_i : i \in [d])$ with eigenvalues in non-decreasing order. Let δ be the vector expressing $(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)$ in the eigenbasis of $\mathbf{H}$. We have

$$F(\boldsymbol{\theta}_k) - F_* = \frac{1}{2}\sum_{i=1}^{d} \delta_i^2 h_i (1 - \gamma h_i)^{2k}. \tag{8}$$

We make assumptions on δ_i and h_i capturing a power-law deceay in the spectrum. In particular, assume

$$h_i \sim \lambda + \frac{L}{i^\alpha} \tag{9}$$

and

$$\delta_i \sim \frac{\Delta}{i^{\beta/2}}\frac{1}{1 + \frac{\lambda}{L}i^\alpha}. \tag{10}$$

Notice under this assumption that $||\delta||_2^2 = ||\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*||_2^2$ is bounded when $\alpha + \beta > 1$ and $\lambda > 0$.

This is similar to the assumptions on the input and target that I discussed in my previous presentation on theoretical attempts to explain scaling laws in random features regression. The parameter λ in the spectrum of h_i arises by assuming we have an additional ridge term $\frac{\lambda}{2}||\theta||_2^2$ in F.

The key to the analysis is consider the collective behaviour of the eigenvalues as opposed to bounding the largest or smallest one. Let

$$a_k := F(\boldsymbol{\theta}_k) - F_* \tag{11}$$

and assume $\gamma \le 1/L$. Consider the regime $d \to \infty$, $k \to \infty$, $\lambda \to 0$ such that $\lambda k \to c$. Then, letting $\kappa = \alpha + \beta$,

$$
\begin{aligned}
a_k &= \frac{1}{2}\sum_{i=1}^{\infty} \delta_i^2 h_i (1-\gamma h_i)^{2k} \\
&\sim \frac{1}{2}\sum_{i=1}^{\infty} \frac{\Delta^2}{i^\beta} \frac{1}{(1+\frac{\lambda}{L} i^\alpha)^2} (\lambda + L i^{-\alpha}) \big(1 - \gamma(\lambda + L i^{-\alpha})\big)^{2k} \\
&= \frac{L\Delta^2}{2}\sum_{i=1}^{\infty} \frac{1}{i^{\kappa-\alpha}} \frac{1}{(1+\frac{\lambda}{L} i^\alpha)^2} i^{-\alpha} \big(1+\tfrac{\lambda}{L} i^\alpha\big) \big(1 - \gamma(\lambda + L i^{-\alpha})\big)^{2k} \\
&= \frac{L\Delta^2}{2}(1-\gamma\lambda)^{2k}\sum_{i=1}^{\infty} \frac{1}{i^\kappa} \frac{1}{(1+\frac{\lambda}{L} i^\alpha)} \left(1 - \frac{\gamma L}{1-\gamma\lambda}\frac{1}{i^\alpha}\right)^{2k}.
\end{aligned}
\tag{12}
$$

Now, letting $\nu = \frac{\lambda k}{L}$, Bach considers the "integral equivalent"

$$
\alpha_k \sim \frac{L\Delta^2}{2}\left(1 - \gamma\frac{L\nu}{k}\right)^{2k} \int_1^\infty \frac{1}{t^\kappa} \frac{1}{1+\frac{\nu}{k}t^\alpha}\left(1 - \frac{\gamma L}{t^\alpha}\right)^{2k} dt. \tag{13}
$$

Now, Bach makes the substitution $u = \frac{2k\gamma L}{t^\alpha}$ and considers "exponential equivalents". In particular,

$$
\left(1 - \gamma\frac{L\nu}{k}\right)^{2k} \sim e^{-2\gamma L\nu}, \tag{14}
$$

$$
\left(1 - \frac{\gamma L}{t^\alpha}\right)^{2k} = \left(1 - \frac{u}{2k}\right)^{2k} \sim e^{-u}. \tag{15}
$$

Moreover,

$$
du = -\frac{2ak\gamma L}{t^{\alpha+1}}\,dt. \tag{16}
$$

Hence,

$$
\alpha_k \sim \frac{L\Delta^2}{2\alpha} e^{-2\nu\gamma L} \frac{1}{(2k\gamma L)^{\frac{\beta-1}{\alpha}+1}} \int_0^1 \frac{u}{u + 2\gamma L\nu} u^{\frac{\beta-1}{\alpha}} e^{-u}\,du. \tag{17}
$$

Indeed, note that

$$
u^{\frac{\beta-1}{\alpha}+1} = \frac{(2\gamma L)^{\frac{\beta-1}{\alpha}+1}}{t^{\alpha+\beta-1}}, \tag{18}
$$

and

$$
\begin{aligned}
t^{\alpha+\beta-1}(u + 2\gamma L\nu) &= 2k\gamma L t^{\beta-1} + 2\gamma L\nu t^{\alpha+\beta-1} \\
&= 2\gamma L k t^{\beta-1}(1 + \tfrac{\nu}{k} t^\alpha) \\
&= 2\gamma L k t^{-\alpha-1} t^\kappa (1 + \tfrac{\nu}{k} t^\alpha) \\
&= -\frac{du}{a} t^\kappa (1 + \frac{\nu}{k} t^\alpha).
\end{aligned}
\tag{19}
$$

Then, using the *exponential integral function*, which satisfies

$$\int_0^\infty \frac{e^{-u}u^{\omega-1}}{u+z}\,du = e^z E_\omega(z)\Gamma(\omega), \tag{20}$$

we have

$$\alpha_k \sim \frac{L\Delta^2}{2\alpha}\frac{\Gamma(\frac{\beta-1}{\alpha}+1)}{(2k\gamma L)^{\frac{\beta-1}{\alpha}+1}}\underbrace{\left(\frac{\beta-1}{\alpha}+1\right)E_{\frac{\beta-1}{\alpha}+2}(2\gamma L\nu)}_{:=c(2\nu\gamma L)}. \tag{21}$$

When $z := 2\nu\gamma L \to 0$, we have $c(z) \to 1$ (power-law convergence). By contrast, when $z \to \infty$, then $c(z) \to (\frac{\beta-1}{\alpha}+1)\frac{e^{-2z}}{z}$ (exponential convergence).

Random feature models. How does this insight apply to GD for random feature models? Consider

$$\begin{aligned} F(\boldsymbol{\theta}) &= \frac{1}{2n}\sum_{i=1}^n \big(y_i - \boldsymbol{\theta}^T\phi(\mathbf{x}_i)\big)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 \\ &= \frac{1}{2n}||\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2. \end{aligned} \tag{22}$$

Here, the Hessian has the form

$$\mathbf{H} = \frac{1}{n}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}, \tag{23}$$

i.e., the sample covariance of the transformed inputs plus the term arising from ridge. Then,

$$\boldsymbol{\theta}_* = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + n\lambda\mathbf{I})^{-1}\boldsymbol{\Phi}^T\mathbf{y}. \tag{24}$$

Assume the random features map ϕ is such that

$$\phi(\mathbf{x})_i = \frac{1}{\sqrt{d}}\psi(\mathbf{x},\mathbf{v}_i) \in \mathbb{R} \tag{25}$$

for $\mathbf{v}_i$ random (we will consider the uniform distribution on $\mathbb{S}^{d-1}$).

Then, by SLLN,

$$\sum_{i=1}^d \phi(\mathbf{x})_i\phi(\mathbf{x}')_i = \frac{1}{d}\sum_{i=1}^d \psi(\mathbf{x},\mathbf{v}_i)\psi(\mathbf{x}',\mathbf{v}_i) \to k(\mathbf{x},\mathbf{x}') \tag{26}$$

for some kernel k. Taking ψ to be the s-th power of the ReLU activation function, it is well known (e.g., Cho and Saul (2009)) that

$$k(x,x') = \frac{1}{\pi}||\mathbf{x}||\,||\mathbf{x}'||\big(\sin\theta + (\pi-\theta)\cos\theta\big), \tag{27}$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{x}'$.

Asymptotically, the eigenvalues of the Gram matrix $\boldsymbol{\Phi}\boldsymbol{\Phi}^T$ are n times the eigenvalues of the integral operator

$$T_k f := \int_{\mathcal{X}} f(\mathbf{x})k(\cdot,\mathbf{x})\,\mu(d\mathbf{x}), \tag{28}$$

where μ is the uniform measure on the $\mathbb{S}^{d-1}$. By Mercer's Theorem, there exist eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ and eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ of this operator such that

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \mu_j \phi_j(\mathbf{x}) \phi_j(\mathbf{x}'). \tag{29}$$

It is shown in Bach (2017) that in our case, $\mu_j \sim j^{-\alpha}$ with $\alpha = 1 + \frac{2s+1}{m-1}$ and $\{\phi_j\}_{j=1}^{\infty}$ a basis of spherical harmonics. Hence, in the asymptotic regime $d, n \to \infty$ (i.e., we approach the limiting kernel), we have

$$h_i \sim \lambda + \frac{L}{i^{\alpha}}. \tag{30}$$

Now, Bach assumes that $y_i = f(x_i)$ is sampled from a Gaussian process with covariance kernel $q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e., $\mathbf{y} \in \mathbb{R}^n$ is Gaussian with covariance being the associated Gram matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$. Assume that k and q have the same eigenfunctions, so that for sufficiently large n, the largest eigenvectors of $\mathbf{K}$ and $\mathbf{Q}$ are the same. This means that $\mathbf{K}$ and $\mathbf{Q}$ "approximately" commute. (This can be checked for our ReLU example). Moreover, we assume that $\mathbf{Q}$ satisfies spectral power law decay with eigenvalues asymptotically of the form $\frac{nM}{i^{\kappa}}$. Recall, we already established that $\mathbf{K}$ has spectrum decaying as $\frac{nL}{i^{\alpha}}$.

Then, using $\boldsymbol{\theta}_* = (\boldsymbol{\Phi}^T \boldsymbol{\Phi} + n\lambda \mathbf{I})^{-1} \boldsymbol{\Phi}^T \mathbf{y}$,

$$\begin{aligned} \delta_i &\sim \frac{\sqrt{nL/i^{\alpha}}}{nL/i^{\alpha} + n\lambda} \frac{n^{1/2} M^{1/2}}{i^{\kappa/2}} z_i \\ &\sim \frac{M^{1/2}/L^{1/2}}{i^{(\kappa-\alpha)/2}} \frac{1}{1 + \frac{\lambda}{L} i^{\alpha}} z_i. \end{aligned} \tag{31}$$

where $z_i \sim \mathcal{N}(0, 1)$. This is of a form that is amenable to the analysis by Laplace's method that we discussed earlier.

References

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53.

Cho, Y. and Saul, L. (2009). Kernel methods for deep learning. *Advances in neural information processing systems*, 22.