Scribe Notes on Spectral Descent

Lecturer: Ying Qi Wen
Scribes: Inzaghi and Yin

1 Motivation: optimization over matrices

A standard first-order optimization problem has the form

$$\min_{x \in V} f(x),$$

where V is a finite-dimensional vector space together with a chosen geometry, typically encoded by a norm. In the more familiar Euclidean setting $V = \mathbb{R}^d$ with $\|\cdot\|_2$, gradient descent with stepsize η updates according to

$$x_{t+1} = x_t - \eta \nabla f(x_t).$$

The key point is that first-order methods depend not only on the vector space structure, but also on the *choice of norm*. Even when parameters can be represented as vectors, Euclidean geometry is only one possible choice. In many applications, such as neural networks, the parameters are naturally organized as matrices, and flattening them into long vectors can obscure important structure such as singular values, rank, and layerwise interactions.

A simple neural network

Consider a feedforward network with N layers. Let

$$h_0 = z, \qquad h_i = \sigma(W_i h_{i-1}), \quad i = 1, \dots, N,$$

where each W_i is a weight matrix and σ is an activation function applied elementwise. The network output is

$$F_W(z) = h_N, \qquad W = (W_1, \dots, W_N).$$

Given a loss function $\mathcal{L}$, we want to solve

$$\min_{W_1, \dots, W_N} \mathcal{L}(F_W).$$

So the search space is no longer $\mathbb{R}^d$ but a **product of matrix spaces:**

$$\mathbb{R}^{d_1 \times d_0} \times \mathbb{R}^{d_2 \times d_1} \times \cdots \times \mathbb{R}^{d_N \times d_{N-1}}.$$

Why not simply vectorize everything?

In principle, one may stack all entries of all weight matrices into a single gigantic vector and then run ordinary gradient descent. Equivalently, one may imagine forming a block variable such as

$$\widetilde{W} = \begin{bmatrix} W_1 & & \\ & W_2 & \\ & & \ddots \end{bmatrix} \qquad \text{or} \qquad \widehat{W} = \begin{bmatrix} -W_1- \\ -W_2- \\ \vdots \end{bmatrix}$$

This is mathematically valid, but it **hides the matrix structure** of the problem. A more natural viewpoint is ***blockwise*** or *layerwise* optimization: update each matrix W_i using its own matrix gradient $\nabla_{W_i}\mathcal{L}$. This leads us to the generic matrix optimization problem

$$\min_{X\in\mathbb{R}^{m\times n}} f(X).$$

The point of spectral descent is that, once the variable is a matrix, we are free to use the space with **norms that are genuinely matrix-aware**, rather than automatically falling back on the Euclidean geometry obtained by vectorization.

Remark 1.1. *The **Frobenius** norm turns $\mathbb{R}^{m\times n}$ into an isometric copy of $\mathbb{R}^{mn}$:*

$$(\mathbb{R}^{m\times n}, \|\cdot\|_F) \cong (\mathbb{R}^{mn}, \|\cdot\|_2).$$

Thus, ordinary gradient descent on matrices is essentially just gradient descent on the vectorized parameter. Spectral descent is different: it uses the operator-norm *geometry of matrices.*

2 Matrix norms and duality

We consider $\mathbb{R}^{m\times n}$ as an inner product space under the Frobenius inner product:

$$\langle A, B\rangle := \mathrm{Tr}(A^\top B).$$

The three most important norms for us are:

- **Operator norm / spectral norm** of A: the largest singular value of A.
$$\|A\|_{op} := \sigma_{\max}(A);$$
- **Frobenius norm**:
$$\|A\|_F := \Big(\sum_{i,j} A_{ij}^2\Big)^{1/2} = \Big(\sum_k \sigma_k(A)^2\Big)^{1/2};$$
- **Nuclear norm / trace norm** of A: the sum of the singular values of A.
$$\|A\|_* := \sum_k \sigma_k(A).$$

If $\sigma(A) = (\sigma_1(A), \ldots, \sigma_r(A))$ denotes the vector of nonzero singular values of A, where $r = \mathrm{rank}(A)$, then

$$\|A\|_{op} = \|\sigma(A)\|_\infty, \qquad \|A\|_F = \|\sigma(A)\|_2, \qquad \|A\|_* = \|\sigma(A)\|_1. \tag{1}$$

Thus the relations among these matrix norms are exactly the familiar relations among ℓ_∞, ℓ_2, and ℓ_1 norms. Therefore,

$$\|A\|_{op} \le \|A\|_F \le \|A\|_*.$$

Moreover, if $r = \mathrm{rank}(A)$, then these norms are equivalent up to factors depending on r:

$$\|A\|_F \le \sqrt{r}\,\|A\|_{op}, \qquad \|A\|_* \le \sqrt{r}\,\|A\|_F, \qquad \|A\|_* \le r\,\|A\|_{op}.$$

Equivalently,

$$\|A\|_{op} \ge \frac{1}{\sqrt{r}}\|A\|_F, \qquad \|A\|_F \ge \frac{1}{\sqrt{r}}\|A\|_*, \qquad \|A\|_{op} \ge \frac{1}{r}\|A\|_*.$$

Dual norms

The Frobenius norm is **self-dual**:

$$\|G\|_F = \max_{\|U\|_F \le 1} \langle G, U \rangle .$$

In contrast, the dual of the operator norm is the nuclear norm:

$$\|G\|_* = \max_{\|U\|_{op} \le 1} \langle G, U \rangle .$$

This is one of the key reasons spectral descent looks different from ordinary gradient descent. These mirror the familiar vector dualities

$$\|g\|_2 = \max_{\|u\|_2 \le 1} \langle g, u \rangle \qquad \text{and} \qquad \|g\|_1 = \max_{\|u\|_\infty \le 1} \langle g, u \rangle.$$

From Equation (1), we see the relationship between the Frobenius norm with ℓ_2, and Nuclear norm with the ℓ_1 norm.

By Von Neumann's trace inequality, maximizing the matrix inner product $\langle G, U \rangle$ reduces to maximizing the vector inner product of their singular values $\langle \sigma(G), \sigma(U) \rangle$. So, the matrix duality:

$$\|G\|_* = \max_{\|U\|_{op} \le 1} \langle G, U \rangle$$

is the singular-value analogue of

$$\|g\|_1 = \max_{\|u\|_\infty \le 1} \langle g, u \rangle.$$

Remark 2.1. *Because $\|A\|_{op} \le \|A\|_F$, the operator-norm unit ball contains the Frobenius unit ball. Dually, the corresponding dual norm is larger:*

$$\|G\|_* \ge \|G\|_F .$$

Equivalently, the nuclear-norm unit ball is contained in the Frobenius unit ball.

Polar factor

For $G \neq 0$, let

$$G = U \Sigma V^\top$$

be a compact singular value decomposition. We define its *polar factor* by

$$\mathrm{polar}(G) := UV^\top.$$

For convenience, define

$$\mathrm{polar}(0) := 0.$$

Then

$$\|\mathrm{polar}(G)\|_{op} \le 1,$$

with equality whenever $G \neq 0$, and

$$\langle G, \mathrm{polar}(G) \rangle = \|G\|_* .$$

Thus, $\mathrm{polar}(G)$ is a **maximizer** in the dual characterization of the nuclear norm:

$$\mathrm{polar}(G) \in \arg \max_{\|U\|_{op} \le 1} \langle G, U \rangle .$$

3 Smoothness in spectral geometry

For a differentiable function on a normed vector space $(V, \|\cdot\|)$, the natural smoothness condition is

$$\|\nabla f(X) - \nabla f(Y)\|_{dual} \leq L \|X - Y\|,$$

where $\|\cdot\|_{dual}$ is the dual norm. When $V = \mathbb{R}^{m\times n}$ is equipped with the Frobenius norm, this becomes the usual Euclidean smoothness condition over matrices:

$$\|\nabla f(X) - \nabla f(Y)\|_F \leq L_F \|X - Y\|_F.$$

For spectral descent, we instead equip the matrix space with the **operator norm**, so the corresponding dual norm on gradients is the nuclear norm.

Definition 3.1 (Spectral smoothness)**.** *A differentiable function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is called L_*-*smooth with respect to the operator norm *if*

$$\|\nabla f(X) - \nabla f(Y)\|_* \leq L_* \|X - Y\|_{op} \qquad \textit{for all } X, Y \in \mathbb{R}^{m\times n}.$$

Equivalently,

$$L_* := \sup_{X\neq Y} \frac{\|\nabla f(X) - \nabla f(Y)\|_*}{\|X - Y\|_{op}}.$$

This is the exact analogue of the standard smoothness condition, but with the primal norm $\|\cdot\|_{op}$ and the dual norm $\|\cdot\|_*$.

Spectral descent lemma

Proposition 3.2 (Descent lemma in operator-norm geometry)**.** *Suppose f is L_*-smooth with respect to $\|\cdot\|_{op}$. Then for all $X, Y \in \mathbb{R}^{m\times n}$,*

$$f(Y) \leq f(X) + \langle \nabla f(X), Y - X\rangle + \frac{L_*}{2}\|Y - X\|_{op}^2.$$

Proof. Let $\Delta := Y - X$. By the fundamental theorem of calculus,

$$\begin{aligned}
f(X+\Delta) - f(X) &= \int_0^1 \langle \nabla f(X + \lambda\Delta), \Delta\rangle d\lambda \\
&= \int_0^1 \langle \nabla f(X) + \nabla f(X + \lambda\Delta) - \nabla f(X), \Delta\rangle d\lambda && \text{(Add/subtract } \nabla f(X)) \\
&= \int_0^1 \langle \nabla f(X), \Delta\rangle d\lambda + \int_0^1 \langle \nabla f(X + \lambda\Delta) - \nabla f(X), \Delta\rangle d\lambda && \text{(Linearity)} \\
&= \langle \nabla f(X), \Delta\rangle + \int_0^1 \langle \nabla f(X + \lambda\Delta) - \nabla f(X), \Delta\rangle d\lambda && \text{(First term constant w.r.t. } \lambda) \\
&\leq \langle \nabla f(X), \Delta\rangle + \int_0^1 \|\nabla f(X + \lambda\Delta) - \nabla f(X)\|_* \|\Delta\|_{op} d\lambda && \text{(Duality: } \langle G, \Delta\rangle \leq \|G\|_*\|\Delta\|_{op}) \\
&\leq \langle \nabla f(X), \Delta\rangle + \int_0^1 L_* \lambda \|\Delta\|_{op}^2 d\lambda && (L_*\text{-smoothness}) \\
&= \langle \nabla f(X), \Delta\rangle + L_*\|\Delta\|_{op}^2 \left[\frac{\lambda^2}{2}\right]_0^1 && \text{(Evaluate integral)} \\
&= \langle \nabla f(X), \Delta\rangle + \frac{L_*}{2}\|\Delta\|_{op}^2.
\end{aligned}$$

□

4 Derivation of spectral descent

Let X_t be the current iterate and write

$$G_t := \nabla f(X_t).$$

By the spectral descent lemma, a natural local upper model of f near X_t, up to an additive constant, is

$$Q_t(\Delta) = \langle G_t, \Delta \rangle + \frac{L_*}{2} \|\Delta\|_{op}^2 .$$

A *spectral descent step* chooses

$$\Delta_t^{SD} \in \arg\min_{\Delta \in \mathbb{R}^{m \times n}} Q_t(\Delta), \qquad X_{t+1} = X_t + \Delta_t^{SD}.$$

Solving the subproblem

We now solve

$$\min_{\Delta} \left\{ \langle G_t, \Delta \rangle + \frac{L_*}{2} \|\Delta\|_{op}^2 \right\}.$$

Write $\Delta = \alpha U$, where $\alpha = \|\Delta\|_{op} \geq 0$ and $\|U\|_{op} \leq 1$. Then

$$\langle G_t, \Delta \rangle + \frac{L_*}{2} \|\Delta\|_{op}^2 = \alpha \langle G_t, U \rangle + \frac{L_*}{2} \alpha^2.$$

For fixed α, the best choice of U is the one that minimizes the linear term, namely

$$U \in \arg\min_{\|U\|_{op} \leq 1} \langle G_t, U \rangle .$$

Using duality,

$$\min_{\|U\|_{op} \leq 1} \langle G_t, U \rangle = - \|G_t\|_*$$

and one minimizer is

$$U = - \operatorname{polar}(G_t).$$

Therefore, the subproblem reduces to

$$\min_{\alpha \geq 0} \left\{ -\alpha \|G_t\|_* + \frac{L_*}{2} \alpha^2 \right\}.$$

The minimizer is

$$\alpha_t^* = \frac{\|G_t\|_*}{L_*}.$$

Hence,

$$\Delta_t^{SD} = -\frac{\|G_t\|_*}{L_*} \operatorname{polar}(G_t).$$

We have proved the following.

Proposition 4.1 (Spectral descent update)**.** *If f is L_*-smooth with respect to $\|\cdot\|_{op}$, then a minimizer of the quadratic model*

$$\Delta \mapsto \langle \nabla f(X_t), \Delta \rangle + \frac{L_*}{2} \|\Delta\|_{op}^2$$

is

$$\Delta_t^{SD} = -\frac{\|\nabla f(X_t)\|_*}{L_*} \operatorname{polar}(\nabla f(X_t)).$$

Therefore, the spectral descent update is

$$X_{t+1} = X_t - \frac{\|\nabla f(X_t)\|_*}{L_*} \operatorname{polar}(\nabla f(X_t)).$$

Remark 4.2 (Hidden **Frank-Wolfe** step)**.** *The direction*

$$D_t \in \arg\min_{\|D\|_{op} \le 1} \langle \nabla f(X_t), D \rangle$$

is exactly the linear minimization oracle over the operator-norm ball, which is the key primitive in Frank–Wolfe methods. Thus spectral descent can be interpreted as follows:

(i) compute the Frank–Wolfe direction over the operator-norm ball;

(ii) scale it optimally using the quadratic model.

5 One-step decrease guarantee

Substituting the spectral descent step into the descent lemma gives a clean decrease bound.

Proposition 5.1 (Guaranteed decrease)**.** *If*

$$X_{t+1} = X_t - \frac{\|\nabla f(X_t)\|_*}{L_*} \operatorname{polar}(\nabla f(X_t)),$$

then

$$f(X_{t+1}) - f(X_t) \le -\frac{1}{2L_*} \|\nabla f(X_t)\|_*^2 .$$

Proof. Let $G_t = \nabla f(X_t)$ and

$$\Delta_t = -\frac{\|G_t\|_*}{L_*} \operatorname{polar}(G_t).$$

By the spectral descent lemma,

$$f(X_t + \Delta_t) - f(X_t) \le \langle G_t, \Delta_t \rangle + \frac{L_*}{2} \|\Delta_t\|_{op}^2 .$$

Since $\langle G_t, \operatorname{polar}(G_t) \rangle = \|G_t\|_*$ and $\|\operatorname{polar}(G_t)\|_{op} \le 1$,

$$\langle G_t, \Delta_t \rangle = -\frac{\|G_t\|_*}{L_*} \langle G_t, \operatorname{polar}(G_t) \rangle = -\frac{\|G_t\|_*^2}{L_*},$$

and

$$\|\Delta_t\|_{op}^2 = \frac{\|G_t\|_*^2}{L_*^2} \|\operatorname{polar}(G_t)\|_{op}^2 \le \frac{\|G_t\|_*^2}{L_*^2}.$$

Therefore

$$f(X_{t+1}) - f(X_t) \le -\frac{\|G_t\|_*^2}{L_*} + \frac{L_*}{2} \frac{\|G_t\|_*^2}{L_*^2} = -\frac{1}{2L_*} \|G_t\|_*^2 .$$

□

Corollary 5.2 (Nonconvex stationarity bound)**.** *Assume f is bounded below by $f_{\inf}$. Then, for the spectral descent update above,*

$$\sum_{t=0}^{T-1} \|\nabla f(X_t)\|_*^2 \le 2L_*\big(f(X_0) - f_{\inf}\big).$$

In particular,

$$\min_{0\le t<T} \|\nabla f(X_t)\|_*^2 \le \frac{2L_*\big(f(X_0) - f_{\inf}\big)}{T}.$$

6 Comparison with ordinary gradient descent

If we instead use Frobenius geometry, then the relevant smoothness constant is

$$L_F := \sup_{X\neq Y} \frac{\|\nabla f(X) - \nabla f(Y)\|_F}{\|X - Y\|_F}.$$

The usual gradient descent update is

$$X_{t+1}^{GD} = X_t - \frac{1}{L_F}\nabla f(X_t),$$

and the standard decrease bound is

$$f(X_{t+1}^{GD}) - f(X_t) \le -\frac{1}{2L_F}\|\nabla f(X_t)\|_F^2.$$

So the two methods give the following guaranteed one-step decreases:

$$\text{Spectral descent:} \qquad -\frac{1}{2L_*}\|\nabla f(X_t)\|_*^2$$

$$\text{Gradient descent:} \qquad -\frac{1}{2L_F}\|\nabla f(X_t)\|_F^2$$

Remark 6.1. *Because $\|G\|_* \ge \|G\|_F$, spectral descent uses a potentially larger gradient measure. On the other hand, one always has $L_* \ge L_F$ as well. Therefore neither method* ***uniformly dominates*** *the other.*

The meaningful comparison is

$$\frac{\|\nabla f(X_t)\|_*^2}{L_*} \quad \textit{versus} \quad \frac{\|\nabla f(X_t)\|_F^2}{L_F}.$$

Which geometry is better depends on the problem and on the structure of the current gradient, such as its rank.

7 A quadratic toy example

To make these constants concrete, consider the matrix least-squares problem

$$f(X) = \frac{1}{2}\|AX - B\|_F^2, \qquad A, X, B \in \mathbb{R}^{d\times d}.$$

Then

$$\nabla f(X) = A^\top(AX - B) = A^\top A\, X - A^\top B.$$

The Hessian is the linear map

$$\nabla^2 f[Z] = A^\top A\, Z.$$

Frobenius smoothness constant

Let $M := A^\top A$. Then

$$L_F = \sup_{Z \neq 0} \frac{\|MZ\|_F}{\|Z\|_F} = \|M\|_{op} = \|A\|_{op}^2 .$$

Spectral smoothness constant

Similarly,

$$L_* = \sup_{Z \neq 0} \frac{\|MZ\|_*}{\|Z\|_{op}}.$$

Since

$$\|MZ\|_* \leq \|M\|_* \|Z\|_{op} ,$$

we have $L_* \leq \|M\|_*$. Taking $Z = I$ gives equality, hence

$$L_* = \|M\|_* = \left\|A^\top A\right\|_* .$$

Because $A^\top A$ is positive semidefinite,

$$\left\|A^\top A\right\|_* = \mathrm{Tr}(A^\top A) = \|A\|_F^2 .$$

Therefore,

$$L_* = \|A\|_F^2 .$$

If we write the singular value decomposition of A as

$$A = P\Sigma Q^\top,$$

then

$$A^\top A = Q\Sigma^2 Q^\top,$$

so indeed

$$\mathrm{Tr}(A^\top A) = \sum_i \sigma_i(A)^2 = \|A\|_F^2 .$$

Comparing the two decrease bounds

For this quadratic problem, the guarantees become

$$f(X_{t+1}^{SD}) - f(X_t) \leq -\frac{1}{2\|A\|_F^2} \|\nabla f(X_t)\|_*^2$$

and

$$f(X_{t+1}^{GD}) - f(X_t) \leq -\frac{1}{2\|A\|_{op}^2} \|\nabla f(X_t)\|_F^2 .$$

Hence the spectral descent bound is stronger than the gradient descent bound exactly when

$$\frac{\|\nabla f(X_t)\|_*^2}{\|\nabla f(X_t)\|_F^2} \geq \frac{\|A\|_F^2}{\|A\|_{op}^2}.$$

The right-hand side is the ***stable rank*** of A:

$$\mathrm{sr}(A) := \frac{\|A\|_F^2}{\|A\|_{op}^2}.$$

So, in this example, spectral descent looks favorable when the gradient is sufficiently **spread out** in singular directions compared with the stable rank of the data matrix A.

Remark 7.1. *This comparison is only a* one-step worst-case bound*. It should not be interpreted as a universal dominance statement. In practice, the overall trajectory, conditioning, stochasticity, and implementation details all matter.*

8 Back to neural networks

Returning to the neural network setting, each layer weight W_i is itself a matrix. A natural blockwise spectral descent update is

$$W_i^{t+1} = W_i^t - \frac{\|\nabla_{W_i}\mathcal{L}(W_t)\|_*}{L_{*,i}}\ \mathrm{polar}\big(\nabla_{W_i}\mathcal{L}(W_t)\big), \qquad i = 1, \dots, N,$$

where $W_t = (W_1^t, \dots, W_N^t)$ and $L_{*,i}$ is a layerwise spectral smoothness constant.

This should be compared with ordinary layerwise gradient descent:

$$W_i^{t+1} = W_i^t - \frac{1}{L_{F,i}}\nabla_{W_i}\mathcal{L}(W_t).$$

Remark 8.1 (On hyperparameter transfer)**.** *One practical advantage of this viewpoint is that it preserves the familiar first-order optimization template: each block is updated using its own gradient and a scalar stepsize. Thus many ideas from standard GD/SGD—learning-rate schedules, per-layer tuning, stochastic gradients, and so on—transfer naturally to the matrix setting, while still respecting matrix geometry.*