2 Gradient Descent

Goal

Linear regression, gradient descent, convergence of gradient, convergence of function value, convergence rate, line search.

Alert 2.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.
We remind that $\langle \cdot, \cdot \rangle$ is the inner product defined in Lecture 0, and $\|\mathbf{w}\|_2 := \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}$.

Definition 2.2: Problem

In this lecture we consider the unconstrained smooth minimization problem

$$f_\star = \inf_{\mathbf{w}} \ f(\mathbf{w}), \tag{2.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and there is no constraint on $\mathbf{w}$. We will consider both convex and nonconvex f.

Example 2.3: Linear least-square regression

Given a dataset consisting of n pairs of feature vectors $X = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and responses $\mathbf{y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$, we are interested in fitting a linear function

$$\hat{y} = \langle \mathbf{x}, \mathbf{w} \rangle + b = \langle \mathbf{x}, \mathbf{w} \rangle, \quad \text{where} \quad \mathbf{x} := \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}, \ \mathbf{w} := \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}.$$

Once $\mathbf{w}$ is estimated, we may then predict the response of an *unseen* feature vector that is not in our dataset.

Our given dataset $X, \mathbf{y}$ may not be generated by a linear function, so we resort to minimizing some fitness function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$:

$$\inf_{\mathbf{w}} \ \frac{1}{n} \sum_{i=1}^n \ell(\hat{y}_i; y_i), \quad \text{where} \quad \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle .$$

For definiteness, choosing $\ell(\hat{y}; y) = (\hat{y} - y)^2$ yields the standard linear least-square regression that goes back at least to Gauss (and Legendre, see here):

$$\min_{\mathbf{w}} \ \frac{1}{n} \sum_{i=1}^n (\langle \mathbf{x}_i, \mathbf{w} \rangle - y_i)^2 \quad = \quad \min_{\mathbf{w}} \ \underbrace{\frac{1}{n} \|\mathsf{X}^\top \mathbf{w} - \mathbf{y}\|_2^2}_{f(\mathbf{w})}, \quad \text{where} \quad \mathsf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n].$$

It is apparent that the above function f is continuously differentiable and our minimization problem is unconstrained. This note contains more about linear regression.

Algorithm 2.4: Gradient descent

Specializing Algorithm 0.41 of feasible direction we obtain the gradient descent algorithm for solving our problem in (2.1):

Algorithm: Gradient descent for unconstrained smooth minimization

Input: $\mathbf{w}_0$

for $t = 0, 1, \ldots$ **do**
- compute gradient $\nabla f(\mathbf{w}_t)$
- choose step size $\eta_t > 0$
- $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \cdot \nabla f(\mathbf{w}_t)$ `// update`

Intuitively, Algorithm 2.4 moves the iterate $\mathbf{w}_t$ along the gradient direction $\nabla f(\mathbf{w}_t)$, which can be justified by Taylor's expansion:

$$f(\mathbf{w}_{t+1}) = f(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)) = f(\mathbf{w}_t) - \eta_t \|\nabla f(\mathbf{w}_t)\|_2^2 + o(\eta_t).$$

Clearly, for small (positive) η_t, $f(\mathbf{w}_{t+1}) < f(\mathbf{w}_t)$, i.e. the algorithm is descending hence making progress. However, descending alone is not guarantee of any meaningful convergence, yet!

In this lecture, we will mostly be concerned with the following choices of the step size η_t:

- Cauchy's rule (Cauchy 1847), where the existence of the minimizer is assumed:

$$\eta_t \in \operatorname*{argmin}_{\eta \geq 0} f(\mathbf{w}_t - \eta \nabla f(\mathbf{w}_t)). \tag{2.2}$$

- Curry's rule (Curry 1944), where the finiteness of η_t is assumed:

$$\eta_t = \inf\{\eta \geq 0 : f'(\mathbf{w}_t - \eta \nabla f(\mathbf{w}_t)) = 0\}.$$

- Constant rule: $\eta_t \equiv \eta > 0$. This is the most practical choice and we will see how to set the constant.

Cauchy, M. A.-L. (1847). "Méthode générale pour la résolution des systémes d'équations simultanées". *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, vol. 25, no. 2, pp. 536–538.

Curry, H. B. (1944). "The Method of Steepest Descent for Non-linear Minimization Problems". *Quarterly of Applied Mathematics*, vol. 2, no. 3, pp. 258–261.

History 2.5: Augustin-Louis Cauchy

The notions of limit and continuity that you learned in Calculus are due to Cauchy, who largely modernized and rigorized calculus. Cauchy's calculus ideas were apparently so cutting-edge that École Polytechnique (where Cauchy taught) had to place note takers in Cauchy's classroom, to make sure he stick to the official course outline (traditional infinitesimal) rather than the kind of rigor that Cauchy desires to teach (to first year engineering students). His famous textbook "Cours d'Analyse" is now available here. Lemaréchal (2012) gave some nice perspective on Cauchy's original work on gradient descent. See also Barany (2013).

While Cauchy deserves the credit for discovering gradient descent, we should also mention Kantorovich (1945), who analyzed the convergence almost at the same time as Curry (even in a more abstract, infinite dimensional setting).

Lemaréchal, C. (2012). "Cauchy and the gradient method". *Documenta Mathematica*, vol. Extra ISMP, pp. 251–254.

Barany, M. J. (2013). "Stuck in the Middle: Cauchy's Intermediate Value Theorem and the History of Analytic Rigor". *Notices of the AMS*, vol. 60, no. 10, pp. 1334–1338.

Kantorovich, L. V. (1945). "On an effective method of solving extremal problems for quadratic functionals". *Soviet Mathematics Doklady*, vol. 48, no. 7, pp. 595–600.

Example 2.6: Descending alone does not guarantee convergence to stationary point

Consider the function

$$f(x)=\begin{cases}\frac{3}{4}(1-x)^2-2(1-x), & x>1\\ \frac{3}{4}(1+x)^2-2(1+x), & x<-1\\ x^2-1, & x\in[-1,1]\end{cases},\ \text{with gradient } f'(x)=\begin{cases}\frac{3}{2}x+\frac{1}{2}, & x>1\\ \frac{3}{2}x-\frac{1}{2}, & x<-1\\ 2x, & x\in[-1,1]\end{cases}.$$

Clearly, f is convex and has a unique minimizer $x_\star=0$ (with $f_\star=-1$). It is easy to verify that $f(x)<f(y)\iff|x|<|y|$. Start with $x_0>1$, set the step size $\eta\equiv 1$, and choose $d=f'(x)$, then $x_1=x_0-f'(x_0)=-\frac{x_0+1}{2}$. By induction it is easy to show that $x_{t+1}=-\frac{1}{2}(x_t-(-1)^t)$. Thus, $x_t>1$ if t is odd and $x_t<-1$ if t is even. Moreover, $|x_{t+1}|<|x_t|$, implying $f(x_{t+1})<f(x_t)$. It is easy to show that $|x_t|\to 1$ and $f(x_t)\to 0$, hence the algorithm is not converging to a stationary point.

Proposition 2.7: A simple and cool result

Let f be a lower semi-continuous (l.s.c.) function on a topological space V, $\mathsf{P}:\mathsf{V}\to\mathsf{V}$ with $f(\mathsf{P}\mathbf{w})\le f(\mathbf{w})$ for all $\mathbf{w}$, and $f(\mathsf{P}\mathbf{w})$ is u.s.c. in $\mathbf{w}$. Then, if the sequence $\mathbf{w}_{t+1}=\mathsf{P}\mathbf{w}_t$ has a limit point $\mathbf{w}_*$, $f(\mathsf{P}\mathbf{w}_*)=f(\mathbf{w}_*)$.

Proof: Clearly the sequence $f(\mathbf{w}_t)$ is monotonically decreasing. Since f is l.s.c. and $f(\mathsf{P}\mathbf{w})$ is u.s.c., we have $f(\mathbf{w}_t)\ge f(\mathbf{w}_*)\ge f(\mathsf{P}\mathbf{w}_*)\ge\limsup_{\mathbf{w}_{t_k}\to\mathbf{w}_*} f(\mathsf{P}\mathbf{w}_{t_k})=\limsup_{\mathbf{w}_{t_k}\to\mathbf{w}_*} f(\mathbf{w}_{t_k+1})\ge\liminf_{\mathbf{w}_{t_k}\to\mathbf{w}_*} f(\mathbf{w}_{t_k+1})\ge f(\mathbf{w}_*)$. ■

Theorem 2.8: Convergence of Cauchy's gradient descent with compactness (Polyak 1963)

Suppose the sublevel set $[\![f\le f(\mathbf{w}_0)]\!]$ is compact, then the sequence $\{\mathbf{w}_t\}$ generated by gradient descent Algorithm 2.4 with Cauchy's step size rule (2.2) satisfies $f(\mathbf{w}_t)\downarrow f_*$ for some $f_*\in\mathbb{R}$ and $\nabla f(\mathbf{w}_t)\to\mathbf{0}$. If f is also convex, then $f(\mathbf{w}_t)\downarrow f_\star$.

Proof: If $\nabla f(\mathbf{w}_k)=0$ for some k, then for all $t\ge k$, $\mathbf{w}_t=\mathbf{w}_k$ and $\nabla f(\mathbf{w}_t)=\mathbf{0}$. On the other hand, if $\nabla f(\mathbf{w}_t)\ne 0$, then

$$f(\mathbf{w}_{t+1})=\min_{\eta\ge 0} f(\mathbf{w}_t-\eta\nabla f(\mathbf{w}_t)):=f(\mathsf{P}\mathbf{w}_t)<f(\mathbf{w}_t).$$

Clearly, $f(\mathsf{P}\mathbf{w})$ is u.s.c. (since ∇f is continuous). Since the set $[\![f\le f(\mathbf{w}_0)]\!]$ is compact, the sequence $\{\mathbf{w}_t\}$ has limit points. By Proposition 2.7 any limit point $\mathbf{w}_*$ must satisfy $f(\mathsf{P}\mathbf{w}_*)=f(\mathbf{w}_*)$, which can happen iff $\nabla f(\mathbf{w}_*)=0$. Now for any subsequence of $\nabla f(\mathbf{w}_t)$ we can extract a further subsequence so that $\mathbf{w}_{t_k}$ converges to say $\mathbf{w}_*$. Thanks to the continuity of ∇f our proof is complete.

If f is convex, then for all $\mathbf{w}$ we have $f(\mathbf{w})\ge f(\mathbf{w}_t)+\langle\mathbf{w}-\mathbf{w}_t,\nabla f(\mathbf{w}_t)\rangle\to f(\mathbf{w}_t)$ (thanks to the compactness of the sublevel set). ■

Note that we have not proved that the iterate $\mathbf{w}_t$ itself will converge, even in the presence of compact sublevel sets!

Polyak, B. T. (1963). "Gradient methods for the minimization of functionals". *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 643–653.

Theorem 2.9: Lipschitz continuity = bounded derivative

Let $\mathsf{T}:(\mathbb{R}^d,\|\cdot\|_{(1)})\to(\mathbb{R}^m,\|\cdot\|_{(2)})$ be differentiable. Then, T is $\mathsf{L}^{[0]}$-Lipschitz continuous, i.e., for all $\mathbf{w}$

and $\mathbf{z}$

$$\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})\|_{(2)} \le \mathsf{L}^{[0]} \cdot \|\mathbf{z}\|_{(1)} \tag{2.3}$$

if and only if for all $\mathbf{w}$

$$\|\mathsf{T}'(\mathbf{w})\| := \sup_{\|\mathbf{z}\|_{(1)}\le 1} \|\mathsf{T}'(\mathbf{w})(\mathbf{z})\|_{(2)} \le \mathsf{L}^{[0]}, \tag{2.4}$$

where T' *is the derivative (gradient) of* T.

Proof: $\Longleftarrow$: Suppose T is $\mathsf{L}^{[0]}$-Lipschitz continuous. By definition of the derivative (see Definition 0.14 in Lecture 0):

$$\begin{aligned} 0 = \lim_{\mathbf{0}\neq\mathbf{z}\to\mathbf{0}} \frac{\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})-\mathsf{T}'(\mathbf{w})(\mathbf{z})\|_{(2)}}{\|\mathbf{z}\|_{(1)}} &\ge \limsup_{\mathbf{0}\neq\mathbf{z}\to\mathbf{0}} \frac{\|\mathsf{T}'(\mathbf{w})(\mathbf{z})\|_{(2)} - \|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})\|_{(2)}}{\|\mathbf{z}\|_{(1)}} \\ &\ge \limsup_{\mathbf{0}\neq\mathbf{z}\to\mathbf{0}} \frac{\|\mathsf{T}'(\mathbf{w})(\mathbf{z})\|_{(2)} - \mathsf{L}^{[0]}\cdot\|\mathbf{z}\|_{(1)}}{\|\mathbf{z}\|_{(1)}}. \end{aligned}$$

Using homogeneity we must have $\|\mathsf{T}'(\mathbf{w})\| \le \mathsf{L}^{[0]}$.

$\Longrightarrow$: Similarly, using the definition of derivative again:

$$\begin{aligned} 0 = \lim_{\mathbf{0}\neq\mathbf{z}\to\mathbf{0}} \frac{\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})-\mathsf{T}'(\mathbf{w})(\mathbf{z})\|_{(2)}}{\|\mathbf{z}\|_{(1)}} &\ge \limsup_{\mathbf{0}\neq\mathbf{z}\to\mathbf{0}} \frac{\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})\|_{(2)} - \|\mathsf{T}'(\mathbf{w})(\mathbf{z})\|_{(2)}}{\|\mathbf{z}\|_{(1)}} \\ &\ge \limsup_{\mathbf{0}\neq\mathbf{z}\to\mathbf{0}} \frac{\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})\|_{(2)} - \mathsf{L}^{[0]}\cdot\|\mathbf{z}\|_{(1)}}{\|\mathbf{z}\|_{(1)}}. \end{aligned}$$

Therefore, for all $\epsilon > 0$ there exists some $\delta > 0$ such that if $\|\mathbf{z}\|_{(1)} \le \delta$ then $\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})\|_{(2)} \le (\mathsf{L}^{[0]}+\epsilon)\|\mathbf{z}\|_{(1)}$. Using triangle inequality we can extend to any $\mathbf{z}$:

$$\|\mathsf{T}(\mathbf{w}+\mathbf{z})-\mathsf{T}(\mathbf{w})\|_{(2)} \le \sum_{i=1}^{n} \|\mathsf{T}(\mathbf{w}+\mathbf{z}_i)-\mathsf{T}(\mathbf{w}+\mathbf{z}_{i-1})\|_2 \le \sum_{i=1}^{n} (\mathsf{L}^{[0]}+\epsilon)\|\mathbf{z}_i-\mathbf{z}_{i-1}\|_{(1)} = (\mathsf{L}^{[0]}+\epsilon)\|\mathbf{z}\|_{(1)},$$

where $\mathbf{z}_i$'s are equally spaced on the line segment $[\mathbf{0},\mathbf{z}]$. Since ϵ is arbitrary, we conclude that T is $\mathsf{L}^{[0]}$-Lipschitz continuous. ■

We have chosen to present the proof in a way that is easy to "relativize." In particular, we can restrict $\mathbf{z}$, the "perturbation" to a cone. Besides, the only if part only requires (2.3) to hold at $\mathbf{w}$ while the if part requires (2.4) to hold on the line segment connecting $\mathbf{w}$ and $\mathbf{z}$.

Theorem 2.10: Taylor's theorem

Let $\mathbf{f}:\mathbb{R}^d \to \mathbb{R}^m$ *be* $(k+1)$*-times continuously differentiable, then*

$$\mathbf{f}(\mathbf{z}) = \mathbf{f}^{[k]}(\mathbf{z};\mathbf{w}) + \int_0^1 \frac{(1-t)^k}{k!} \mathbf{f}^{(k+1)}\big(\mathbf{w}+t(\mathbf{z}-\mathbf{w})\big)(\mathbf{z}-\mathbf{w})^{\otimes(k+1)}\,\mathrm{d}t, \quad \textit{where}$$

$$\mathbf{f}^{[k]}(\mathbf{z};\mathbf{w}) = \sum_{\ell=0}^{k} \frac{1}{\ell!} f^{(\ell)}(\mathbf{w})(\mathbf{z}-\mathbf{w})^{\otimes\ell},$$

and $\mathbf{f}^{(\ell)}$ *is the* ℓ*-th order derivative and* $\mathbf{u}^{\otimes\ell} := \underbrace{\mathbf{u}\otimes\cdots\otimes\mathbf{u}}_{\ell \text{ times}}$ *is the* ℓ*-fold tensor product.*

Proof: The result is well-known for $m = 1$. For $m > 1$, consider applying the theorem to real-valued functions $f_{\mathbf{u}} := \langle \mathbf{f}; \mathbf{u} \rangle$ for all $\mathbf{u} \in \mathbb{R}^m$, or see Cartan (1971, Theorem 5.6.1) for a direct proof. ■

We remind that the expansion $f^{[k]}(\mathbf{z}; \mathbf{w})$ is a polynomial in $\mathbf{z}$!

Cartan, H. (1971). "Differential Calculus". Hermann.

Definition 2.11: $\mathsf{L}^{[k]}$-smoothness

We call a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ $\mathsf{L}^{[k]}$-smooth w.r.t. the norm $\|\cdot\|$ on $\mathbb{R}^d$ if the following one-sided inequality holds:

$$\forall \mathbf{w}, \forall \mathbf{z}, \quad f(\mathbf{z}) - f^{[k]}(\mathbf{z}; \mathbf{w}) \le \frac{\mathsf{L}^{[k]}}{(k+1)!} \|\mathbf{z} - \mathbf{w}\|^{k+1}, \tag{2.5}$$

i.e. the function f is bounded by its k-th Taylor expansion plus some constant multiple of the $(k+1)$-th power of the norm. Clearly, the (smallest) constant $\mathsf{L}^{[k]}$ depends on the norm $\|\cdot\|$.

If $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ is vector-valued, we need to scalarize the left-hand side of (2.5). We do so through a support function σ_C on the range space $\mathbb{R}^m$:

$$\sigma_C(\mathbf{s}) = \sup_{\mathbf{c} \in C} \langle \mathbf{c}; \mathbf{s} \rangle,$$

where the compact set $C \subseteq \mathbb{R}^m$ can be taken w.l.o.g. to be convex (or the extreme points thereof). We easily verify that σ_C is finite-valued and sublinear (hence also continuous). We call $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ $(\mathsf{L}^{[k]}, C)$-smooth w.r.t. the norm $\|\cdot\|$ on $\mathbb{R}^d$ and the set $C \subseteq \mathbb{R}^m$ if

$$\forall \mathbf{w}, \mathbf{z}, \quad \sigma_C\big(\mathbf{f}(\mathbf{z}) - \mathbf{f}^{[k]}(\mathbf{z}; \mathbf{w})\big) \le \frac{\mathsf{L}^{[k]}}{(k+1)!} \|\mathbf{z} - \mathbf{w}\|^{k+1},$$

or equivalently, the real-valued functions $\{\langle \mathbf{c}; \mathbf{f} \rangle : \mathbf{c} \in C\}$ are all $\mathsf{L}^{[k]}$-smooth.

Proposition 2.12: Equivalence of $(\mathsf{L}^{[k]}, C)$-smoothness

For a vector-valued smooth function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$, we have (II) $\implies$ (III) $\iff$ (I) $\implies$ (IV):

(I). $\mathbf{f}$ *is* $(\mathsf{L}^{[k]}, C)$*-smooth, i.e.* $\forall \mathbf{w}, \forall \mathbf{z}, \ \sigma_C\big(\mathbf{f}(\mathbf{z}) - \mathbf{f}^{[k]}(\mathbf{z}; \mathbf{w})\big) \le \frac{\mathsf{L}^{[k]}}{(k+1)!} \|\mathbf{z} - \mathbf{w}\|^{k+1}$.

(II). $\forall \mathbf{w}, \forall \mathbf{z}, \ \sigma_C\big([\mathbf{f}^{(k)}(\mathbf{z}) - \mathbf{f}^{(k)}(\mathbf{w})](\mathbf{z} - \mathbf{w})^{\otimes k}\big) \le \mathsf{L}^{[k]} \|\mathbf{z} - \mathbf{w}\|^{k+1}$.

(III). $\forall \mathbf{w}, \forall \mathbf{z}, \ \sigma_C\big(\mathbf{f}^{(k+1)}(\mathbf{w}) \mathbf{z}^{\otimes (k+1)}\big) \le \mathsf{L}^{[k]} \|\mathbf{z}\|^{k+1}$.

(IV). $\forall \mathbf{w}, \forall \mathbf{z}, \ \sigma_C\big(\mathbf{f}(\mathbf{z}) - \mathbf{f}^{[k]}(\mathbf{z}; \mathbf{w}) + \mathbf{f}(\mathbf{w}) - \mathbf{f}^{[k]}(\mathbf{w}; \mathbf{z})\big) \le \frac{2\mathsf{L}^{[k]}}{(k+1)!} \|\mathbf{z} - \mathbf{w}\|^{k+1}$.

When $k = 1$, all four are equivalent.

Proof: It suffices to prove the real-valued case. (I) $\implies$ (IV): swapping $\mathbf{w}$ and $\mathbf{z}$ in item (I) we have

$$f(\mathbf{w}) - f^{[k]}(\mathbf{w}; \mathbf{z}) \le \frac{\mathsf{L}^{[k]}}{(k+1)!} \|\mathbf{z} - \mathbf{w}\|^{k+1}.$$

Adding the two inequalities then yields item (IV).

(II) $\implies$ (III): Dividing both sides of item (II) by $\|\mathbf{z} - \mathbf{w}\|^{k+1}$ and letting $\frac{\mathbf{z} - \mathbf{w}}{\|\mathbf{z} - \mathbf{w}\|} \to \mathbf{u}$ we obtain

$$f^{(k+1)}(\mathbf{w}) \mathbf{u}^{\otimes (k+1)} \le \mathsf{L}^{[k]},$$

which is exactly item (III).

(III) $\implies$ (I): Applying the Taylor Theorem 2.10 we have

$$\begin{aligned} f(\mathbf{z}) - f^{(k)}(\mathbf{z};\mathbf{w}) &= \int_0^1 \frac{(1-t)^k}{k!} f^{(k+1)}\big(\mathbf{w}+t(\mathbf{z}-\mathbf{w})\big)(\mathbf{z}-\mathbf{w})^{\otimes(k+1)}\,\mathrm{d}t \\ &\le \int_0^1 \frac{(1-t)^k}{k!}\mathsf{L}^{[k]}\|\mathbf{z}-\mathbf{w}\|^{k+1}\,\mathrm{d}t = \frac{\mathsf{L}^{[k]}}{(k+1)!}\|\mathbf{z}-\mathbf{w}\|^{k+1}. \end{aligned}$$

(I) $\implies$ (III): Dividing both sides of item (I) by $\|\mathbf{z}-\mathbf{w}\|^{k+1}$ and letting $\frac{\mathbf{z}-\mathbf{w}}{\|\mathbf{z}-\mathbf{w}\|} \to \mathbf{u}$. ■

Sometimes it is more convenient to rewrite (III) as

$$\|\mathbf{f}^{(k+1)}(\mathbf{w})\|_C := \sup_{\|\mathbf{u}\|=1} \sigma_C\big(\mathbf{f}^{(k+1)}(\mathbf{w})\mathbf{u}^{\otimes(k+1)}\big) \le \mathsf{L}^{[k]},$$

i.e. the semi-norm of the symmetric tensor $\mathbf{f}^{(k+1)}(\mathbf{w})$ is uniformly bounded by $\mathsf{L}^{[k]}$.

Theorem 2.13: Lipschitz continuity implies $\mathsf{L}^{[k]}$-smoothness

For a vector-valued smooth function $\mathbf{f}:\mathbb{R}^d \to \mathbb{R}^m$, we have (I) $\iff$ (II) $\implies$ (III):

(I). $\mathbf{f}^{(k)}$ is $\mathsf{L}^{[k]}$-Lipschitz continuous, i.e.

$$\forall \mathbf{w},\forall \mathbf{z},\ \|\mathbf{f}^{(k)}(\mathbf{w}) - \mathbf{f}^{(k)}(\mathbf{z})\|_C \le \mathsf{L}^{[k]}\|\mathbf{w}-\mathbf{z}\|, \quad \text{where} \quad \|\mathcal{A}\|_C := \sup_{\|\mathbf{u}\|=1} \sigma_C\big(\mathcal{A}(\mathbf{u}^{\otimes k})\big).$$

(II). $\mathbf{f}^{(k+1)}$ is $\mathsf{L}^{[k]}$-bounded, i.e.

$$\forall \mathbf{w},\ \|\mathbf{f}^{(k+1)}(\mathbf{w})\|_C^\circ := \sup_{\|\mathbf{v}\|=1} \|\mathbf{f}^{(k+1)}(\mathbf{w})\mathbf{v}\|_C = \sup_{\|\mathbf{v}\|=1}\sup_{\|\mathbf{u}\|=1} \sigma_C\big(\mathbf{f}^{(k+1)}(\mathbf{w})(\mathbf{v}\otimes\mathbf{u}^{\otimes k})\big) \le \mathsf{L}^{[k]}.$$

(III). $\mathbf{f}$ is $(\mathsf{L}^{[k]}, C)$-smooth, where for odd k, we may replace C with its symmetric hull $C \cup -C$.

Moreover, for $k=1$, all three are equivalent if either of the following holds:

- *if $\langle \mathbf{c};\mathbf{f}\rangle$ is convex or concave for all $\mathbf{c} \in C$ or*
- *the norm $\|\cdot\|$ is Euclidean (where C is w.l.o.g. symmetric).*

Proof: The equivalence of (I) and (II) follows from Theorem 2.9 (with $\mathsf{T} = \mathbf{f}^{(k)}$, $\|\cdot\|_{(1)} = \|\cdot\|$ and $\|\cdot\|_{(2)} = \|\cdot\|_C$), while the implication to (III) follows from Proposition 2.12. Note that for odd k, $\|\mathcal{A}\|_C = \|\mathcal{A}\|_{C\cup -C}$.

From now on let $k=1$. We first verify (III) $\implies$ (II) when f is convex. Indeed, from convexity we have

$$0 \le f''(\mathbf{w})(\mathbf{v}-\mathbf{u})\otimes(\mathbf{v}-\mathbf{u}) = f''(\mathbf{w})\mathbf{v}\otimes\mathbf{v} - f''(\mathbf{w})\mathbf{v}\otimes\mathbf{u} - f''(\mathbf{w})\mathbf{u}\otimes\mathbf{v} + f''(\mathbf{w})\mathbf{u}\otimes\mathbf{u}.$$

Using symmetry, we have

$$2f''(\mathbf{w})\mathbf{u}\otimes\mathbf{v} \le f''(\mathbf{w})\mathbf{v}\otimes\mathbf{v} + f''(\mathbf{w})\mathbf{u}\otimes\mathbf{u} \le 2\sup_{\|\mathbf{u}\|=1} f''(\mathbf{w})\mathbf{u}\otimes\mathbf{u}.$$

Since $\mathbf{u}$ and $\mathbf{v}$ are arbitrary unit vectors, from Proposition 2.12 it follows (II) hence also (I).

When the norm is Euclidean, we have the surprising equivalence for all functions:

$$\begin{aligned} \forall \mathbf{w},\forall \mathbf{z},\ \|f'(\mathbf{w}) - f'(\mathbf{z})\|_2 \le \mathsf{L}^{[1]}\|\mathbf{w}-\mathbf{z}\|_2 &\iff |[f'(\mathbf{w}) - f'(\mathbf{z})](\mathbf{w}-\mathbf{z})| \le \mathsf{L}^{[1]}\|\mathbf{w}-\mathbf{z}\|_2^2 \\ &\iff \pm f \text{ is } \mathsf{L}^{[1]}\text{-smooth}, \end{aligned}$$

due to the fact that

$$\rho(f''(\mathbf{w})) := \sup_{\|\mathbf{u}\|_2=1} |f''(\mathbf{w})\mathbf{u}\otimes\mathbf{u}| = \sup_{\|\mathbf{u}\|_2=1}\sup_{\|\mathbf{v}\|_2=1} f''(\mathbf{w})\mathbf{u}\otimes\mathbf{v} = \|f''(\mathbf{w})\|_{\mathrm{sp}},$$

i.e. the spectral radius coincides with the spectral norm for a symmetric matrix $f''(\mathbf{w})$. ■

Corollary 2.14: Lipschitz continuous gradient implies $\mathsf{L}^{[1]}$-smoothness

Consider the following statements for a real-valued smooth function:

(I). Vector-valued derivative $f' : \mathbb{R}^d \to \mathbb{R}^d$ is $\mathsf{L}^{[1]}$-Lipschitz continuous, i.e.

$$\forall \mathbf{w}, \forall \mathbf{z},\ \|f'(\mathbf{w}) - f'(\mathbf{z})\|_\circ \le \mathsf{L}^{[1]}\|\mathbf{w} - \mathbf{z}\|.$$

(II). Matrix-valued second-order derivative $f'' : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is $\mathsf{L}^{[1]}$-bounded, i.e.

$$\forall \mathbf{w},\ \sup_{\|\mathbf{u}\|=1} \|f''(\mathbf{w})\mathbf{u}\|_\circ = \sup_{\|\mathbf{u}\|=1}\sup_{\|\mathbf{v}\|=1} \langle f''(\mathbf{w})\mathbf{u}; \mathbf{v}\rangle \le \mathsf{L}^{[1]}.$$

(III). Real-valued functions $\pm f$ are $\mathsf{L}^{[1]}$-smooth, i.e.,

$$\forall \mathbf{w}, \forall \mathbf{z},\ \ |f(\mathbf{z}) - f(\mathbf{w}) - f'(\mathbf{w})(\mathbf{z} - \mathbf{w})| \le \tfrac{\mathsf{L}^{[1]}}{2}\|\mathbf{z} - \mathbf{w}\|^2.$$

(Recall that f is $\mathsf{L}^{[1]}$-smooth if we remove the absolute value on the left-hand side.)

Then, (I) $\iff$ (II) $\implies$ (III). If f is convex or the norm is Euclidean, then all three are equivalent. ■

We remind that $\|\cdot\|_\circ$ is the dual norm of $\|\cdot\|$, see Definition 0.10. For simplicity one may take $\|\cdot\| = \|\cdot\|_\circ = \|\cdot\|_2$, the Euclidean norm. Obviously, the Lipschitz constant $\mathsf{L}^{[1]}$ depends on the norm! The superscript $^{[1]}$ reminds us that the 1st derivative is Lipschitz continuous.

We note that for a convex function f, $-f$ is trivially $\mathsf{L}^{[1]}$-smooth (see Equation (0.4)). We have thus proved the following surprising equivalence for convex functions:

$$\begin{aligned}\forall \mathbf{w}, \forall \mathbf{z},\ \|f'(\mathbf{w}) - f'(\mathbf{z})\|_\circ \le \mathsf{L}^{[1]}\|\mathbf{w} - \mathbf{z}\| &\iff \langle \mathbf{w} - \mathbf{z}; f'(\mathbf{w}) - f'(\mathbf{z})\rangle \le \mathsf{L}^{[1]}\|\mathbf{w} - \mathbf{z}\|^2\\ &\iff f \text{ is } \mathsf{L}^{[1]}\text{-smooth.}\end{aligned}$$

Alert 2.15: Importance of $\mathsf{L}^{[1]}$-smoothness

Let us explain the intuitive appeal of $\mathsf{L}^{[1]}$-smoothness: In applying the gradient descent Algorithm 2.4 we need to decide the step size η_t at each iteration. If the gradient changes quickly, then we must adjust the step size η_t quickly as well (since the update depends on the product $\eta_t \nabla f(\mathbf{w}_t)$). Locally (e.g. when the step size is infinitesimally small), the rate of change of the gradient is an important index for the gradient descent Algorithm 2.4. As shown in Corollary 2.14, the former behavior is bounded by the derivative of the gradient, namely the Hessian.

Another way to look at $\mathsf{L}^{[1]}$-smoothness is through the equivalence (III) in Corollary 2.14: A (convex) function f is sandwiched by its linear approximation, plus a quadratic term that scales with the constant $\mathsf{L}^{[1]}$. Thus, a smaller $\mathsf{L}^{[1]}$ means the function f can be approximated by a "flatter" quadratic residual term $\frac{\mathsf{L}^{[1]}}{2}\|\mathbf{z} - \mathbf{w}\|^2$, and hence a larger step size would probably not lead us astray.

Similarly, for higher order algorithms, such as the Newton's algorithm that relies on the Hessian, $\mathsf{L}^{[2]}$-smoothness or more generally $\mathsf{L}^{[k]}$-smoothness will play the role of the $\mathsf{L}^{[1]}$-smoothness for gradient algorithms. Moreover, $\mathsf{L}^{[0]}$-smoothness is important as well, for zero-order algorithms and for function approximations, as we will see in later lectures.

Example 2.16: Constant step size for linear regression

Recall our objective function f in linear least-square regression (Example 2.3):

$$n \cdot f(\mathbf{w}) = \|\mathsf{X}^\top \mathbf{w} - \mathbf{y}\|_2^2 = (\mathsf{X}^\top \mathbf{w} - \mathbf{y})^\top (\mathsf{X}^\top \mathbf{w} - \mathbf{y}) = \mathbf{w}^\top \mathsf{X}\mathsf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathsf{X}\mathbf{y} + \mathbf{y}^\top \mathbf{y}.$$

Using the Hessian formula from Example 0.16, we have

$$\nabla^2 f(\mathbf{w}) = \frac{2}{n}\mathsf{X}\mathsf{X}^\top,$$

from which we immediately find that

- f is convex since the matrix $\mathsf{X}\mathsf{X}^\top \succeq \mathbf{0}$;
- f is $\mathsf{L}_2^{[1]}$-smooth under the Euclidean norm $\|\cdot\|_2$, where

$$\mathsf{L}_2^{[1]} = \frac{2}{n}\|\mathsf{X}\mathsf{X}^\top\|_{\mathrm{sp}} = \frac{2}{n}\|\mathsf{X}\|_{\mathrm{sp}}^2 \le \frac{2}{n}\|\mathsf{X}\|_{\mathsf{F}}^2,$$

  where $\|\cdot\|_{\mathrm{sp}}$ is the spectral norm (i.e. largest singular value).

Thus, we may use the constant step size $\eta = \frac{1}{\mathsf{L}_2^{[1]}}$ (see Theorem 2.17 below) in gradient descent for linear regression.

The Hessian happens to be a constant matrix for our example here. More generally, we need to take the maximum over the entire space:

$$\mathsf{L}_2^{[1]} = \sup_{\mathbf{w}} \|\nabla^2 f(\mathbf{w})\|_{\mathrm{sp}}.$$

Theorem 2.17: Convergence of gradient descent for $\mathsf{L}^{[1]}$-smooth functions (Polyak 1963)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathsf{L} = \mathsf{L}_2^{[1]}$-smooth (w.r.t. $\|\cdot\|_2$) and bounded from below (i.e. $f_\star > -\infty$). If the step size $\eta_t \in [\alpha, \frac{2}{\mathsf{L}} - \beta]$ for some $\alpha, \beta > 0$, then the sequence $\{\mathbf{w}_t\}$ generated by Algorithm 2.4 satisfies $\nabla f(\mathbf{w}_t) \to \mathbf{0}$. Moreover,

$$\min_{0 \le t \le T-1} \|\nabla f(\mathbf{w}_t)\|_2 \le \sqrt{\frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta\mathsf{L}T/2}}. \tag{2.6}$$

Proof: We first prove a result that will be used repeatedly:

$$f(\mathbf{w} - \eta\nabla f(\mathbf{w})) \le f(\mathbf{w}) - \eta\|\nabla f(\mathbf{w})\|_2^2 + \frac{\mathsf{L}\eta^2}{2}\|\nabla f(\mathbf{w})\|_2^2 = f(\mathbf{w}) - \eta(1 - \tfrac{\mathsf{L}\eta}{2})\|\nabla f(\mathbf{w})\|_2^2. \tag{2.7}$$

Therefore, if $\eta \in]0, \frac{2}{\mathsf{L}}[$ and $\nabla f(\mathbf{w}) \neq \mathbf{0}$, we *strictly* decrease the function value after each iteration of the gradient descent algorithm. Rearranging we have

$$\|\nabla f(\mathbf{w}_t)\|_2^2 \le \frac{f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})}{\eta_t(1 - \mathsf{L}\eta_t/2)} \le \frac{f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})}{\alpha\beta\mathsf{L}/2}.$$

Summing from $t = 0$ to $t = T - 1$:

$$\sum_{t=0}^{T-1} \|\nabla f(\mathbf{w}_t)\|_2^2 \le \frac{f(\mathbf{w}_0) - f(\mathbf{w}_T)}{\alpha\beta\mathsf{L}/2} \le \frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta\mathsf{L}/2}.$$

Therefore, the sequence $\|\nabla f(\mathbf{w}_t)\|_2$ is square summable hence $\nabla f(\mathbf{w}_t) \to \mathbf{0}$ and the bound (2.6) holds. In other words, the *minimal* gradient $\nabla f(\mathbf{w}_t)$ converges to $\mathbf{0}$ at the rate $1/\sqrt{T}$. ■

Of course, we can tune α and β to optimize the bound (2.6): since $\alpha+\beta \le \frac{2}{\mathsf{L}}$, the minimum is achieved when $\alpha=\beta=\frac{1}{\mathsf{L}}$, and the bound reduces to

$$\min_{0\le t\le T-1} \|\nabla f(\mathbf{w}_t)\|_2 \le \sqrt{\frac{2\mathsf{L}[f(\mathbf{w}_0)-f_\star]}{T}}, \tag{2.8}$$

which enjoys several remarkable properties:

- It is proportional to the Lipschitz smoothness L. Indeed, the bigger L is , the smaller the step size $\eta=\frac{1}{\mathsf{L}}$ is since the function f becomes steeper.
- If we start from some point $\mathbf{w}_0$ whose function value is closer to the infimum $f_\star$, then the gradient diminishes faster to zero.
- Very importantly, the rate of convergence does not depend on d, the dimension, at all!
- The $1/\sqrt{T}$ rate of convergence for the gradient is essentially tight (Cartis et al. 2010).

Polyak, B. T. (1963). "Gradient methods for the minimization of functionals". *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 643–653.
Cartis, C., N. I. M. Gould, and P. L. Toint (2010). "On the Complexity of Steepest Descent, Newton's and Regularized Newton's Methods for Nonconvex Unconstrained Optimization". *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2833–2852.

Example 2.18: Iterates of gradient descent may not converge

With $\eta=\frac{2}{\mathsf{L}^{[1]}}$, gradient descent may not converge (although the function value still converges to some value). Simply revisit Example 2.6. Note also that the iterate $\mathbf{w}_t$ may not converge:

- When there is no minimizer at all: simply consider the convex function

$$f(x)=\begin{cases} e^{-x}, & x\le 0 \\ x+1, & x>0 \end{cases}.$$

  For a multivariate example, use $\mathbf{x}\mapsto \frac{1}{1+\|\mathbf{x}\|_2^2}$.

- Even when there are many minimizers: consider the smoothed Mexican hat function (Absil et al. 2005)

$$f(r,\theta):=\begin{cases} e^{-\frac{1}{1-r^2}}\left[1-\frac{4r^4}{4r^4+(1-r^2)^4}\sin\left(\theta-\frac{1}{1-r^2}\right)\right], & \text{if } r<1 \\ 0, & \text{if } r\ge 1 \end{cases},$$

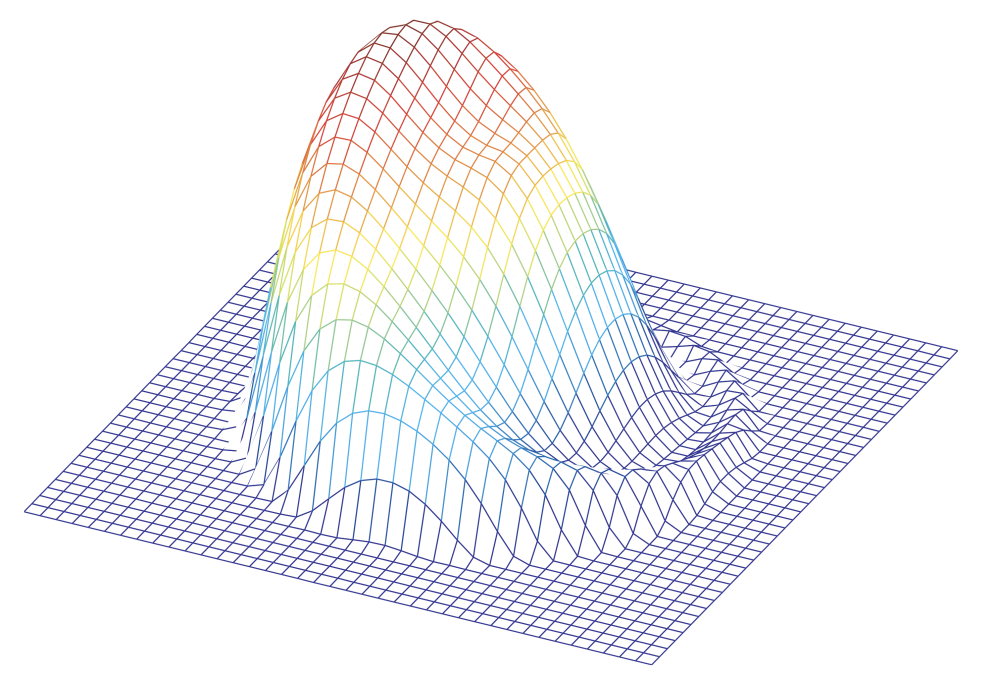
  where (r,θ) denote the polar coordinates in $\mathbb{R}^2$. Observe that $\frac{4r^4}{4r^4+(1-r^2)^4}\in[0,1]$ and hence $f>0$ for all $r<1$. Since sin (and its derivatives) are bounded, the bump function $e^{-\frac{1}{1-r^2}}$ makes sure that all derivatives of f exist at $r=1$ and hence everywhere. Thus, $f\in\mathcal{C}^\infty$ (infinitely many times differentiable) but not analytic. If we initialize with

$$\theta_0(1-r_0^2)=1, \quad r_0<1,$$

  then gradient descent (with infinitesimal step size) will stay on the trajectory

$$\theta(t)[1-r^2(t)]=1,$$

  meaning that $r(t)\to 1$ and hence $\theta(t)$ admits any real number as a limit point.



Absil, P., R. Mahony, and B. Andrews (2005). "Convergence of the Iterates of Descent Methods for Analytic Cost Functions". *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547.

Remark 2.19: Gradient descent minimizes the quadratic upper bound

We may derive gradient descent from minimizing the quadratic upper bound in the definition of $\mathsf{L}_2^{[1]}$-smoothness:

$$f(\mathbf{w}) \leq f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2.$$

Minimizing $\mathbf{w}$ on the right-hand side gives exactly the gradient descent update:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t).$$

This interpretation immediately inspires us to generalize gradient descent to any norm by minimizing a similar upper bound, as will be seen in a later lecture.

Remark 2.20: Backtracking line search (Armijo 1966)

We make two important observations about the proof of Theorem 2.17:

- Since gradient descent is descending, we may restrict $\mathsf{L}^{[1]}$-smoothness to the sublevel set:

$$\mathsf{L}^{[1]} = \sup_{\mathbf{w} \in [\![f \leq f(\mathbf{w}_0)]\!]} \|\nabla^2 f(\mathbf{w})\|.$$

  (What happens outside my vicinity does not matter, as long as I do not step out of my cubical!)

- Computing $\mathsf{L}^{[1]}$ in $\mathsf{L}^{[1]}$-smoothness can be difficult in itself, or simply resulting in step size that is too conservative. Fortunately, as noted by Armijo (1966), the key inequality (2.7) is testable: we start with some trial step size $\eta_0 > 0$, and find the *smallest* $k \in \mathbb{N}$ such that for some fixed $\alpha \in (0, 1)$, e.g.

$\alpha = \frac{1}{2}$, the following key inequality is met:

$$f\big(\mathbf{w} - \eta_k \cdot \mathsf{J}^*(\nabla f(\mathbf{w}))\big) \le f(\mathbf{w}) - \alpha\eta_k \|\nabla f(\mathbf{w})\|_\circ^2, \quad \text{where} \quad \eta_k := \eta_0/2^k. \tag{2.9}$$

Comparing to (2.7) which holds for any η, the above inequality clearly holds when

$$\eta_k \le \frac{2(1-\alpha)}{\mathsf{L}^{[1]}} \iff k \ge K := \left\lceil \log_2 \tfrac{\eta_0 \mathsf{L}^{[1]}}{1-\alpha} \right\rceil - 1.$$

Thus, even without knowing $\mathsf{L} = \mathsf{L}^{[1]}$, by backtracking in (2.9) we can still find a legitimate step size $\eta_t = \eta_0/2^K \ge \frac{1-\alpha}{\mathsf{L}}$. The price to pay is that:

- Our step size might be a constant $1-\alpha$ multiple (or worse if our initial guess η_0 is too small) of the "optimal" one $\frac{1}{\mathsf{L}}$, increasing the bound (2.8) by a constant factor: $\frac{1}{\sqrt{\eta_0\mathsf{L}(2-\eta_0\mathsf{L})}}$ when $\eta_0 \le \frac{2(1-\alpha)}{\mathsf{L}}$ and $\frac{1}{\sqrt{1-\alpha^2}}$ when $\eta_0 > \frac{2(1-\alpha)}{\mathsf{L}}$.
- Backtracking costs (at most) K function evaluations (on the left-hand side of (2.9)).

The potential gain is also huge, since we may avoid using an excessively small step size, especially in the initial phase of the iteration.

Armijo, L. (1966). "Minimization of functions having Lipschitz continuous first partial derivatives". *Pacific Journal of Mathematics*, vol. 16, no. 1, pp. 1–3.

Example 2.21: Gradient descent may converge to saddle point

Consider the function $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$, whose gradient is $\nabla f(x,y) = \binom{x}{y^3-y}$ and Hessian is $\begin{bmatrix} 1 & 0 \\ 0 & 3y^2-1 \end{bmatrix}$. Clearly, there are three stationary points $(0,0), (0,1), (0,-1)$, where the the last two are global minimizers whereas the first is a saddle point. Take any $\mathbf{w}_0 = (x_0, 0)$, then $\mathbf{w}_t = (x_t, 0)$ hence it can only converge to $(0,0)$, which is a saddle point.

Define $\mathcal{W} \subseteq \mathbb{R}^d$ as the set of initializers $\mathbf{w}_0$ such that gradient descent converges to the saddle point $(0,0)$. Lee et al. (2016) proved that $\mathcal{W}$ has measure 0 (in $\mathbb{R}^d$)! In other words, if we randomly initialize $\mathbf{w}_0$, we will almost surely avoid saddle points and converge to local minima (under a so-called strict saddle assumption which is satisfied in our example here).

Lee, J. D., M. Simchowitz, M. I. Jordan, and B. Recht (2016). "Gradient Descent Only Converges to Minimizers". In: *29th Annual Conference on Learning Theory*, pp. 1246–1257.