$$= \frac{2\vec{x}}{1 + \|\vec{x}\|_2^2}. \tag{3.89}$$

For the Hessian, we have

$$\begin{aligned} \nabla^2 h(\vec{x}) &= D(\nabla h)(\vec{x}) && (3.90) \\ &= D\left(\frac{2\vec{x}}{1 + \|\vec{x}\|_2^2}\right). && (3.91) \end{aligned}$$

We compute this Jacobian, hence the desired Hessian, componentwise, and obtain

$$\begin{aligned} [\nabla^2 h(\vec{x})]_{j,k} &= \left[D\left(\frac{2\vec{x}}{1 + \|\vec{x}\|_2^2}\right)\right]_{j,k} && (3.92) \\ &= \frac{\partial}{\partial x_k}\left(\frac{2x_j}{1 + \|\vec{x}\|_2^2}\right) && (3.93) \\ &= 2\frac{(1 + \|\vec{x}\|_2^2)\frac{\partial}{\partial x_k}(x_j) - x_j\frac{\partial}{\partial x_k}(1 + \|\vec{x}\|_2^2)}{(1 + \|\vec{x}\|_2^2)^2} && (3.94) \\ &= 2\frac{(1 + \|\vec{x}\|_2^2)\frac{\partial}{\partial x_k}(x_j) - x_j\frac{\partial}{\partial x_k}(\|\vec{x}\|_2^2)}{(1 + \|\vec{x}\|_2^2)^2} && (3.95) \\ &= 2\frac{(1 + \|\vec{x}\|_2^2)\frac{\partial x_j}{\partial x_k} - 2x_jx_k}{(1 + \|\vec{x}\|_2^2)^2} && (3.96) \\ &= -\frac{4x_jx_k}{(1 + \|\vec{x}\|_2^2)^2} + \frac{2}{1 + \|\vec{x}\|_2^2}\frac{\partial x_j}{\partial x_k} && (3.97) \\ &= -\frac{4x_jx_k}{(1 + \|\vec{x}\|_2^2)^2} + \begin{cases} \frac{2}{1+\|\vec{x}\|_2^2}, & \text{if } j = k \\ 0, & \text{if } j \neq k. \end{cases} && (3.98) \end{aligned}$$

This gives

$$\begin{aligned} [\nabla^2 h(\vec{x})]_{j,k} &= -\frac{4x_jx_k}{(1 + \|\vec{x}\|_2^2)^2}, \qquad \forall j \neq k && (3.99) \\ [\nabla^2 h(\vec{x})]_{jj} &= \frac{2}{1 + \|\vec{x}\|_2^2} - \frac{4x_jx_k}{(1 + \|\vec{x}\|_2^2)^2}, \qquad \forall j. && (3.100) \end{aligned}$$

We can write this using vectors as

$$\nabla^2 h(\vec{x}) = \frac{2}{1 + \|\vec{x}\|_2^2} I - \frac{4\vec{x}\vec{x}^\top}{(1 + \|\vec{x}\|_2^2)^2}. \tag{3.101}$$

3.2 Taylor's Theorems

In this section, we will introduce Taylor approximation and Taylor's theorem for the familiar scalar function case. Then we will generalize the idea of Taylor approximation to multivariate functions. Taylor approximation is a tool to find polynomial approximation of functions using information about the function value at a point along with the value of its firs, second and higher order derivatives.

Definition 67 (Taylor Approximation)

Let $f\colon \mathbb{R} \to \mathbb{R}$ be a k-times continuously differentiable function, and fix $x_0 \in \mathbb{R}$. The k^{th} degree Taylor approximation around x_0 is the function $\widehat{f}_k(\cdot; x_0)\colon \mathbb{R} \to \mathbb{R}$ given by

$$\widehat{f}_k(x; x_0) = f(x_0) + \frac{1}{1!}\frac{\mathrm{d}f}{\mathrm{d}x}(x_0) \cdot (x - x_0) + \cdots + \frac{1}{k!}\frac{\mathrm{d}^k f}{\mathrm{d}x^k}(x_0) \cdot (x - x_0)^k \tag{3.102}$$

$$= \sum_{i=0}^{k} \frac{1}{i!}\frac{\mathrm{d}^i f}{\mathrm{d}x^i}(x_0) \cdot (x - x_0)^i. \tag{3.103}$$

In particular, the first-order and second-order Taylor approximations of f around x_0 are

$$\widehat{f}_1(x; x_0) = f(x_0) + \frac{\mathrm{d}f}{\mathrm{d}x}(x_0) \cdot (x - x_0) \tag{3.104}$$

$$\widehat{f}_2(x; x_0) = f(x_0) + \frac{\mathrm{d}f}{\mathrm{d}x}(x_0) \cdot (x - x_0) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x_0) \cdot (x - x_0)^2. \tag{3.105}$$

We will derive multivariable versions of these approximations later.

Example 68 (Taylor Approximation of Cubic Function)**.** Let us approximate the function $f(x) = x^3$ around the fixed point $x_0 = 1$ using Taylor approximations of different degrees.

$$\widehat{f}_1(x; 1) = f(x_0) + \frac{\mathrm{d}f}{\mathrm{d}x}(x_0) \cdot (x - x_0) \tag{3.106}$$

$$= x_0^3 + 3x_0^2 \cdot (x - x_0) \tag{3.107}$$

$$= 1^3 + 3 \cdot 1^2 \cdot (x - 1) \tag{3.108}$$

$$= 3(x - 1) + 1 \tag{3.109}$$

$$= 3x - 2. \tag{3.110}$$

$$\widehat{f}_2(x; 1) = \widehat{f}_1(x; 1) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x_0) \cdot (x - x_0)^2 \tag{3.111}$$

$$= 3x - 2 + 3 \cdot 1 \cdot (x - 1)^2 \tag{3.112}$$

$$= 3x^2 - 3x + 1. \tag{3.113}$$

$$\widehat{f}_3(x; 1) = \widehat{f}_2(x; 1) + \frac{1}{6}\frac{\mathrm{d}^3 f}{\mathrm{d}x^3}(x_0) \cdot (x - x_0)^3 \tag{3.114}$$

$$= 3x^2 - 3x + 1 + (x - 1)^3 \tag{3.115}$$

$$= x^3. \tag{3.116}$$

We notice the following take-aways:

- The first-order Taylor approximation $\widehat{f}_1(\cdot; x_0)$ is the *best linear approximation* to f around $x = x_0 = 1$. In particular, its graph is the tangent line to the graph of f around the point $(x_0, f(x_0)) = (1, 1)$, as observed in Figure 3.6.

- The second-order Taylor approximation $\widehat{f}_2(\cdot; x_0)$ is the *best quadratic approximation* to f around $x = x_0 = 1$. It is the parabola whose graph passes through the point $(x_0, f(x_0)) = (1, 1)$, as observed in Figure 3.6, and it has the same first and second derivatives as f at x_0. Using the intuition that the second derivative models curvature, we see that the second-order Taylor approximation captures the local curvature of the graph of the function. This intuition will be helpful later when discussing convexity.

- The third-degree Taylor approximation $\widehat{f}_3(\cdot; x_0)$ is the *best cubic approximation* to f; because f is just a cubic function, the best cubic approximation is just f itself, and indeed we have $\widehat{f}_3(\cdot; x_0) = f$.

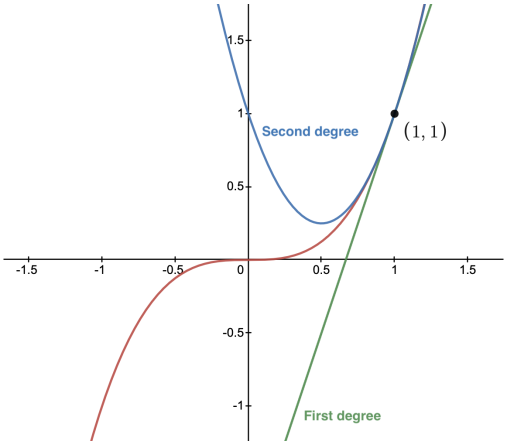

Figure 3.6: First and second degree Taylor approximations of the function $f(x) = x^3$.

Taylor approximation gives us the degree k polynomial that approximates the function $f(x)$ around the fixed point $x = x_0$. Taylor's theorem quantifies the bounds for the error of this approximation.

Theorem 69 (Taylor's Theorem)

Let $f: \mathbb{R} \to \mathbb{R}$ be a function which is k-times continuously differentiable, and fix $x_0 \in \mathbb{R}$. Then for all $x \in \mathbb{R}$ we have

$$f(x) = \widehat{f}_k(x; x_0) + o(|x - x_0|^k) \tag{3.117}$$

where the term $o(|x - x_0|^k)$ (i.e., the *remainder*) denotes a function, say $R_k(x; x_0)$, such that

$$\lim_{x \to x_0} \frac{R_k(x; x_0)}{|x - x_0|^k} = 0. \tag{3.118}$$

We use this remainder notation because we don't really care about what it is precisely, only its limiting behavior as $x \to x_0$, and the little-o notation allows us to not worry too much about the exact form of the remainder.

This theorem certifies that the Taylor approximations $\widehat{f}_k$ are good approximations to f. Another way to write this result is generally more useful or simpler:

$$f(x + \delta) = \underbrace{f(x) + \frac{\mathrm{d}f}{\mathrm{d}x}(x) \cdot \delta}_{=\widehat{f}_1(x+\delta; x)} + o(|\delta|) \tag{3.119}$$

$$= \underbrace{f(x) + \frac{\mathrm{d}f}{\mathrm{d}x}(x) \cdot \delta + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x) \cdot \delta^2}_{=\widehat{f}_2(x+\delta; x)} + o(\delta^2) \tag{3.120}$$

$$= \ldots. \tag{3.121}$$

We will never need to quantitatively work with the remainder in this course; we will usually write $f \approx \widehat{f}_k$ and leave it at that.

3.2.1 Taylor Approximation of Multivariate Functions

Using the definitions we introduced for the gradient and Hessian of multivariate functions, we can generalize the idea of Taylor's approximation to these functions.

Definition 70 (Multivariate Taylor Approximations)

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ and fix $\vec{x}_0 \in \mathbb{R}^n$.

- If f is continuously differentiable, then its first-order Taylor approximation around $\vec{x}_0$ is the function $\widehat{f}_1(\cdot;\vec{x}_0)\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$\widehat{f}_1(\vec{x};\vec{x}_0) = f(\vec{x}_0) + [\nabla f(\vec{x}_0)]^\top(\vec{x} - \vec{x}_0). \tag{3.122}$$

- If f is twice continuously differentiable, then its second-order Taylor approximation around $\vec{x}_0$ is the function $\widehat{f}_2(\cdot;\vec{x}_0)\colon \mathbb{R}^n \to \mathbb{R}$ given by

$$\widehat{f}_2(\vec{x};\vec{x}_0) = f(\vec{x}_0) + [\nabla f(\vec{x}_0)]^\top(\vec{x} - \vec{x}_0) + \frac{1}{2}(\vec{x} - \vec{x}_0)^\top[\nabla^2 f(\vec{x}_0)](\vec{x} - \vec{x}_0). \tag{3.123}$$

The graph of the first-order Taylor approximation is the hyperplane tangent to the graph of f at the point $(\vec{x}_0, f(\vec{x}_0))$. This hyperplane has normal vector $\nabla f(\vec{x}_0)$.

We could define higher-order Taylor approximations $\widehat{f}_k$, but to express them concisely would require generalizations of matrices, called *tensors*. For example, the third derivative of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is a rank-3 tensor, i.e., an object which lives in $\mathbb{R}^{n\times n\times n}$. These are out of scope for this course, and anyways we will only need the first two derivatives.

We can also state an analogous Taylor's theorem.

Theorem 71 (Taylor's Theorem)

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function which is k-times continuously differentiable, and fix $\vec{x}_0 \in \mathbb{R}^n$. Then for all $\vec{x} \in \mathbb{R}^n$ we have

$$f(\vec{x}) = \widehat{f}_k(\vec{x};\vec{x}_0) + o(\|\vec{x} - \vec{x}_0\|_2^k). \tag{3.124}$$

We can re-write this result in the following, more useful, way for $k = 1$ and $k = 2$:

$$\begin{aligned}
f(\vec{x} + \vec{\delta}) &= \underbrace{f(\vec{x}) + [\nabla f(\vec{x})]^\top \vec{\delta}}_{\widehat{f}_1(\vec{x}+\vec{\delta};\vec{x})} + o(\|\vec{\delta}\|_2) && (3.125)\\
&= \underbrace{f(\vec{x}) + [\nabla f(\vec{x})]^\top \vec{\delta} + \frac{1}{2}\vec{\delta}^\top[\nabla^2 f(\vec{x})]\vec{\delta}}_{\widehat{f}_2(\vec{x}+\vec{\delta};\vec{x})} + o(\|\vec{\delta}\|_2^2) && (3.126)\\
&= \dots. && (3.127)
\end{aligned}$$

Example 72 (Taylor Approximation of the Squared ℓ^2 norm)**.** In this example we will compute and visualize the first and second degree Taylor approximations of the squared ℓ^2 norm function $f(\vec{x}) = \|\vec{x}\|_2^2$ for $\vec{x} \in \mathbb{R}^2$ around the vector $\vec{x} = \vec{x}_0$. First recall the gradient and hessian of the function which are computed in Examples 55 and 65, respectively.

- First degree approximation:

$$\begin{aligned}
\widehat{f}_1(\vec{x};\vec{x}_0) &= f(\vec{x}_0) + [\nabla f(\vec{x}_0)]^\top(\vec{x} - \vec{x}_0) && (3.128)\\
&= \|\vec{x}_0\|_2^2 + [2\vec{x}_0]^\top(\vec{x} - \vec{x}_0) && (3.129)\\
&= 2\vec{x}_0^\top \vec{x} - \|\vec{x}_0\|_2^2. && (3.130)
\end{aligned}$$

Recall that the graph of $f(\vec{x})$ has a paraboloid shape. Let us now evaluate and visualize the first degree approximation of the function around $\vec{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

$$\hat{f}_1(\vec{x}; \vec{x}_0) = 2x_1 - 1. \tag{3.131}$$

We plot this function in Figure 3.7 and notice that the graph of the first order approximation is the plane tangent to the paraboloid at the point $(1, 0, f(1, 0)) = (1, 0, 1)$.

- Second degree approximation:

$$\begin{aligned}
\widehat{f}_2(\vec{x}; \vec{x}_0) &= \widehat{f}_1(\vec{x}; \vec{x}_0) + \frac{1}{2}(\vec{x} - \vec{x}_0)^\top [\nabla^2 f(\vec{x}_0)](\vec{x} - \vec{x}_0) && (3.132)\\
&= \underbrace{2\vec{x}_0^\top \vec{x} - \|\vec{x}_0\|_2^2}_{=\widehat{f}_1(\vec{x}; \vec{x}_0)} + \frac{1}{2}(\vec{x} - \vec{x}_0)^\top [2I](\vec{x} - \vec{x}_0) && (3.133)\\
&= 2\vec{x}_0^\top \vec{x} - \vec{x}_0^\top \vec{x}_0 + (\vec{x} - \vec{x}_0)^\top (\vec{x} - \vec{x}_0) && (3.134)\\
&= 2\vec{x}_0^\top \vec{x} - \vec{x}_0^\top \vec{x}_0 + \vec{x}^\top \vec{x} - \vec{x}_0^\top \vec{x} - \vec{x}^\top \vec{x}_0 + \vec{x}_0^\top \vec{x}_0 && (3.135)\\
&= 2\vec{x}_0^\top \vec{x} - \vec{x}_0^\top \vec{x}_0 + \vec{x}^\top \vec{x} - 2\vec{x}_0^\top \vec{x} + \vec{x}_0^\top \vec{x}_0 && (3.136)\\
&= \vec{x}^\top \vec{x} && (3.137)\\
&= \|\vec{x}\|_2^2 . && (3.138)
\end{aligned}$$

Thus $\widehat{f}_2 = f$ independently of the choice of $\vec{x}_0$, which makes sense since f is a quadratic function.

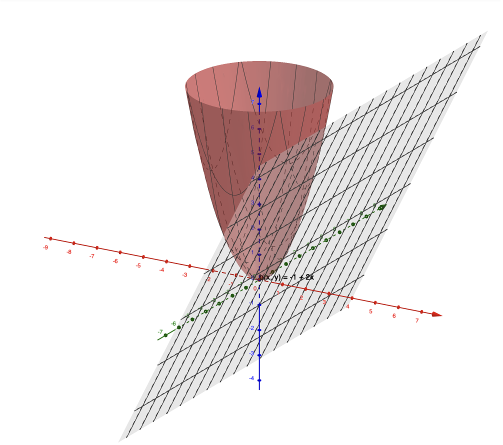

Figure 3.7: First degree Taylor approximation of the function $f(\vec{x}) = \|\vec{x}\|_2^2$

Example 73. We can also compute gradients using Taylor's theorem by pattern matching; this is sometimes much neater than taking componentwise gradients. At first glance this seems circular, but we will see how it is possible. Take for example the function $f: \mathbb{R}^n \to \mathbb{R}$ given by $f(\vec{x}) = \vec{x}^\top A\vec{x}$. We can perturb f around $\vec{x}$ to obtain

$$\begin{aligned}
f(\vec{x} + \vec{\delta}) &= (\vec{x} + \vec{\delta})^\top A(\vec{x} + \vec{\delta}) && (3.139)\\
&= \vec{x}^\top A\vec{x} + \vec{\delta}^\top A\vec{x} + \vec{x}^\top A\vec{\delta} + \vec{\delta}^\top A\vec{\delta} && (3.140)\\
&= f(\vec{x}) + (\vec{x}^\top A^\top + \vec{x}^\top A)\vec{\delta} + \vec{\delta}^\top A\vec{\delta} && (3.141)\\
&= f(\vec{x}) + ((A + A^\top)\vec{x})^\top \vec{\delta} + \frac{1}{2}\vec{\delta}^\top (A + A^\top)\vec{\delta}. && (3.142)
\end{aligned}$$

However, Taylor's theorem tells us that

$$f(\vec{x}+\vec{\delta}) = f(\vec{x}) + [\nabla f(\vec{x})]^\top \vec{\delta} + \frac{1}{2}\vec{\delta}^\top [\nabla^2 f(\vec{x})]\vec{\delta} + o(\|\vec{\delta}\|_2^2). \tag{3.143}$$

By pattern matching, we see that $\nabla f(\vec{x}) = (A + A^\top)\vec{x}$ (as obtained in a previous example), $\nabla^2 f(\vec{x}) = A + A^\top$, and the remainder term is $\vec{0}$.

One final note is that we changed $2A \to A + A^\top$ in Equation (3.142); this is to ensure that the Hessian is symmetric, which is a consequence of Theorem 64; and we are able to do this because

$$\vec{\delta}^\top (2A)\vec{\delta} = \vec{\delta}^\top A\vec{\delta} + \vec{\delta}^\top A\vec{\delta} = \vec{\delta}^\top A\vec{\delta} + (\vec{\delta}^\top A\vec{\delta})^\top = \vec{\delta}^\top A\vec{\delta} + \vec{\delta}^\top A^\top\vec{\delta} = \vec{\delta}^\top (A + A^\top)\vec{\delta}. \tag{3.144}$$

We conclude by introducing a more general version of a first-order Taylor approximation, a corresponding Taylor's theorem, and giving an example of when it is useful.

Definition 74 (Vector-Valued Taylor Approximation)

Let $\vec{f} \colon \mathbb{R}^n \to \mathbb{R}^m$ and fix $\vec{x}_0 \in \mathbb{R}^n$. If $\vec{f}$ is continuously differentiable, then its first-order Taylor approximation around $\vec{x}_0$ is the function $\hat{\vec{f}}_1 \colon \mathbb{R}^n \to \mathbb{R}^m$ given by

$$\hat{\vec{f}}_1(\vec{x}; \vec{x}_0) = \vec{f}(\vec{x}_0) + [D\vec{f}(\vec{x}_0)](\vec{x} - \vec{x}_0). \tag{3.145}$$

Again, higher-order approximations will require higher-order derivatives, which requires tensors.

Theorem 75 (Vector-Valued Taylor's Theorem)

Let $\vec{f} \colon \mathbb{R}^n \to \mathbb{R}^m$ be a continuously differentiable function, and fix $\vec{x}_0 \in \mathbb{R}^n$. Then for all $\vec{x} \in \mathbb{R}^n$ we have

$$\vec{f}(\vec{x}) = \hat{\vec{f}}_1(\vec{x}; \vec{x}_0) + \vec{o}(\|\vec{x} - \vec{x}_0\|_2). \tag{3.146}$$

We can again re-write this result in a more workable form:

$$\vec{f}(\vec{x}+\vec{\delta}) = \vec{f}(\vec{x}) + [D\vec{f}(\vec{x})]\vec{\delta} + o(\|\vec{\delta}\|_2). \tag{3.147}$$

Example 76. Taylor's theorem can be used to compute gradients by pattern matching, even when the function is not linear or quadratic. For instance, we now use it to derive the chain rule (albeit with stronger assumptions on the functions). Let $\vec{f} \colon \mathbb{R}^p \to \mathbb{R}^m$ and $\vec{g} \colon \mathbb{R}^n \to \mathbb{R}^p$ be continuously differentiable. Let $\vec{h} \colon \mathbb{R}^n \to \mathbb{R}^m$ be defined as $\vec{h}(\vec{x}) = \vec{f}(\vec{g}(\vec{x}))$. Then we compute $\vec{h}$ on a perturbation around $\vec{x}$ and expand:

$$\begin{aligned}
\vec{h}(\vec{x}+\vec{\delta}) &= \vec{f}(\vec{g}(\vec{x}+\vec{\delta})) && (3.148)\\
&\approx \vec{f}(\vec{g}(\vec{x}) + [D\vec{g}(\vec{x})]\vec{\delta}) && (3.149)\\
&\approx \vec{f}(\vec{g}(\vec{x})) + [D\vec{f}(\vec{g}(\vec{x}))]([D\vec{g}(\vec{x})]\vec{\delta}) && (3.150)\\
&\approx \vec{f}(\vec{g}(\vec{x})) + [D\vec{f}(\vec{g}(\vec{x}))][D\vec{g}(\vec{x})]\vec{\delta}. && (3.151)
\end{aligned}$$

The first Taylor expansion is an expansion of $\vec{g}$ around the point $\vec{x}$ with perturbation $\vec{\delta}$; the second Taylor expansion is an expansion of $\vec{f}$ around the point $g(\vec{x})$ with perturbation $[D\vec{g}(\vec{x})]\vec{\delta}$.

Meanwhile, Taylor's theorem says that

$$\vec{h}(\vec{x}+\vec{\delta}) \approx \vec{h}(\vec{x}) + [D\vec{h}(\vec{x})]\vec{\delta}. \tag{3.152}$$

Thus by pattern matching we find that

$$D\vec{h}(\vec{x}) = [D\vec{f}(\vec{g}(\vec{x}))][D\vec{g}(\vec{x})] \tag{3.153}$$

which is precisely the chain rule!

Note that here we did not invoke the little-o notation because it turns out to be quite messy, but it is indeed possible to do the required rigorous manipulations and get the same result.

As a last practical note, remembering the formula for Taylor approximations helps us confirm our understanding of the dimensions of each vector. For instance, every term should multiply to a scalar. This makes it simpler to remember that, for a function $f \colon \mathbb{R}^n \to \mathbb{R}$, the gradient ∇f outputs column vectors in $\mathbb{R}^n$, the Hessian $\nabla^2 f$ outputs square matrices in $\mathbb{R}^{n\times n}$, etc.

3.3 The Main Theorem

In this section, we will use the concepts introduced in the previous sections to state and prove one of the fundamental ideas in optimization.

Theorem 77 (The Main Theorem [4])
Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, and let $\Omega \subseteq \mathbb{R}^n$ be an open set.[a] Consider the optimization problem

$$\min_{\vec{x}\in\Omega} f(\vec{x}). \tag{3.154}$$

Let $\vec{x}^\star$ be a solution to this optimization problem. Then

$$\nabla f(\vec{x}^\star) = \vec{0}. \tag{3.155}$$

[a] "Open sets" are analogous to open intervals (a, b) — i.e., not containing boundary points.

This theorem gives a *necessary* condition for a point to be an optimal solution of this optimization problem. It says that any point that is optimal must necessarily have gradient equal to zero.

Proof. We prove this for scalar functions $f \colon \mathbb{R} \to \mathbb{R}$ only; the vector case is a bit more complicated and is left as an exercise.

Using Taylor approximation of the function around the optimal point:

$$f(x) = f(x^\star) + \frac{\mathrm{d}f}{\mathrm{d}x}(x^\star)\cdot(x - x^\star) + o(|x - x^\star|). \tag{3.156}$$

Since $f(x^\star) \leq f(x)$ for all $x \in \Omega$, we have

$$f(x) \leq f(x) + \frac{\mathrm{d}f}{\mathrm{d}x}(x^\star)\cdot(x - x^\star) + o(|x - x^\star|) \tag{3.157}$$

$$\implies 0 \leq \frac{\mathrm{d}f}{\mathrm{d}x}(x^\star)\cdot(x - x^\star) + o(|x - x^\star|). \tag{3.158}$$

Since Ω is an open set, there exists some ball of positive radius $r > 0$ around $x^\star$ such that $B_r(x^\star) \subseteq \Omega$. Formally,

$$B_r(x^\star) = \{x \in \mathbb{R} \mid |x - x^\star| \leq r\}. \tag{3.159}$$

Let us partition $B_r(x^\star)$ into B_+, the set of all $x \in B_r(x^\star)$ such that $x - x^\star \geq 0$, and B_-, the set of all $x \in B_r(x^\star)$ such that $x - x^\star < 0$.

For all $x \in B_{+}$, we have

$$\begin{aligned} 0 &\leq \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) \cdot (x - x^{\star}) + o(|x - x^{\star}|) && (3.160) \\ &= \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) \cdot |x - x^{\star}| + o(|x - x^{\star}|) && (3.161) \\ \implies 0 &\leq \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) + \frac{o(|x - x^{\star}|)}{|x - x^{\star}|}. && (3.162) \end{aligned}$$

Taking the limit as $x \to x^{\star}$ within B_{+}, we have

$$\begin{aligned} 0 &\leq \lim_{\substack{x \to x^{\star} \\ x \in B_{+}}} \left\{ \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) + \frac{o(|x - x^{\star}|)}{|x - x^{\star}|} \right\} && (3.163) \\ &= \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) + \lim_{\substack{x \to x^{\star} \\ x \in B_{+}}} \frac{o(|x - x^{\star}|)}{|x - x^{\star}|} && (3.164) \\ &= \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}). && (3.165) \end{aligned}$$

Thus we have $0 \leq \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star})$. On the other hand, for all $x \in B_{-}$, we have

$$\begin{aligned} 0 &\leq \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) \cdot (x - x^{\star}) + o(|x - x^{\star}|) && (3.166) \\ &= -\frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) \cdot |x - x^{\star}| + o(|x - x^{\star}|) && (3.167) \\ \implies 0 &\geq \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) - \frac{o(|x - x^{\star}|)}{|x - x^{\star}|}. && (3.168) \end{aligned}$$

Taking the limit $x \to x^{\star}$ within B_{-}, we have

$$\begin{aligned} 0 &\geq \lim_{\substack{x \to x^{\star} \\ x \in B_{-}}} \left\{ \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) - \frac{o(|x - x^{\star}|)}{|x - x^{\star}|} \right\} && (3.169) \\ &= \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) - \lim_{\substack{x \to x^{\star} \\ x \in B_{-}}} \frac{o(|x - x^{\star}|)}{|x - x^{\star}|} && (3.170) \\ &= \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}). && (3.171) \end{aligned}$$

Thus we have $0 \geq \frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star})$ and so $\frac{\mathrm{d}f}{\mathrm{d}x}(x^{\star}) = 0$. □

3.4 Directional Derivatives

Recall the definition of the partial derivative of a multivariate function (Definition 47), which represents the rate of change of the function $f(\vec{x})$ along one of the standard basis vectors. We do not need to restrict our treatment to the standard basis vectors; in fact, we can compute the rate of change of the function in any arbitrary direction. This is called the directional derivative.

Definition 78 (Directional Derivative)

Let $f\colon \mathbb{R}^{n} \to \mathbb{R}$ be differentiable, and fix $\vec{u} \in \mathbb{R}^{n}$ such that $\|\vec{u}\|_{2} = 1$. The *directional derivative* of f along $\vec{u}$ is the function $Df(\cdot)[\vec{u}] \colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$Df(\vec{x})[\vec{u}] = \lim_{h \to 0} \frac{f(\vec{x} + h \cdot \vec{u}) - f(\vec{x})}{h}. \tag{3.172}$$

If we know the directional derivative in any direction, we know the gradient; similarly, if we know the gradient, we know the directional derivative. The way to connect the two is given by the following proposition, whose proof is left as an exercise.

Proposition 79

Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be differentiable, and fix $\vec{u} \in \mathbb{R}^n$ such that $\|\vec{u}\|_2 = 1$. Then

$$Df(\vec{x})[\vec{u}] = \vec{u}^\top [\nabla f(\vec{x})]. \tag{3.173}$$

In particular, $Df(\vec{x})[\vec{e}_i] = \frac{\partial f}{\partial x_i}(\vec{x})$.

3.5 (OPTIONAL) Matrix Calculus

So far we have only discussed derivatives of three types of function and all have either scalars or vectors and their input and output. We can think of a more general class of functions that also involve matrices. In this section, we will generalize the idea of derivatives to such functions. We will focus our attention on functions of the form $f \colon \mathbb{R}^{m \times n} \to \mathbb{R}$, which take a matrix $X \in \mathbb{R}^{m \times n}$ as input and produce a scalar $f(X)$ as output. Familiar examples of such functions include matrix norms, the determinant, and the trace.

Definition 80 (Gradient)

Let $f \colon \mathbb{R}^{m \times n} \to \mathbb{R}$ be differentiable. The *gradient* of f is the function $\nabla f \colon \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ which is defined as

$$\nabla f(X) = \begin{bmatrix} \frac{\partial f}{\partial X_{11}}(X) & \cdots & \frac{\partial f}{\partial X_{1n}}(X) \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{m1}}(X) & \cdots & \frac{\partial f}{\partial X_{mn}}(X) \end{bmatrix} \tag{3.174}$$

There exists a general chain rule for matrix-valued functions, which is provable by flattening out all matrices into vectors and applying the vector chain rule.

Theorem 81 (Chain Rule)

Let $F \colon \mathbb{R}^{p \times q} \to \mathbb{R}^{r \times s}$ and $G \colon \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times q}$ be differentiable functions. Let $H \colon \mathbb{R}^{m \times n} \to \mathbb{R}^{r \times s}$ be defined by $H(X) = F(G(X))$ for all $X \in \mathbb{R}^{m \times n}$. Then H is differentiable, and for all i, j, k, ℓ, we have

$$\frac{\partial H_{ij}}{\partial X_{k\ell}}(X) = \sum_a \sum_b \frac{\partial F_{ij}}{\partial G_{ab}}(G(X)) \frac{\partial G_{ab}}{\partial X_{k\ell}}(X) \tag{3.175}$$

As before, the notation $\frac{\partial F_{ij}}{\partial G_{ab}}$ means to take the derivative of the ij^{th} output of F by its ab^{th} input. A more specific version of this chain rule is given below for functions $f \colon \mathbb{R}^{m \times n} \to \mathbb{R}$.

Proposition 82

Let $F\colon \mathbb{R}^{p\times q} \to \mathbb{R}$ and $G\colon \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ be differentiable functions. Let $h\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ be defined by $h(X) = f(G(X))$ for all $X \in \mathbb{R}^{m\times n}$. Then h is differentiable, and for all k, ℓ, we have

$$\frac{\partial h}{\partial X_{k\ell}}(X) = \sum_a \sum_b [\nabla f(G(X))]_{ab} \frac{\partial G_{ab}}{\partial X_{k\ell}}(X) \tag{3.176}$$

We also are able to define a first-order Taylor expansion without having to use tensor notation.

Definition 83 (Matrix Taylor Approximation)

Let $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ and fix $X_0 \in \mathbb{R}^{m\times n}$. If f is continuously differentiable, then its first-order Taylor approximation around X_0 is the function $\widehat{f}_1(\cdot; X_0)\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ given by

$$\widehat{f}_1(X; X_0) = f(X_0) + \operatorname{tr}\left([\nabla f(X_0)]^\top (X - X_0)\right). \tag{3.177}$$

There is a corresponding Taylor's theorem certifying the Taylor approximation accuracy, but we don't state it here.

Finally, note that the general recipe for computing all quantities such as the gradient, Jacobian, and gradient matrix is the same: consider each input component and each output component separately and organize their partial derivatives in vector or matrix form with a standard layout.

Example 84 (Finishing Example 62)**.** Now that we know how to take matrix-valued gradients, we complete the example of Neural Networks and Backpropagation. Before reading the following, please revise the lengthy setup of this example.

We promised in this example a way to compute $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, or more precisely a way to compute $\nabla_{W^{(i)}} L(\boldsymbol{\theta})$. We now have the tools to do this using the chain rule. Recall that we have access to $\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})$ by backpropagation. Then we can compute the components of $\nabla_{W^{(i)}} L(\boldsymbol{\theta})$ by

$$\begin{aligned}
[\nabla_{W^{(i)}} L(\boldsymbol{\theta})]_{j,k} &= \frac{\partial L}{\partial (W^{(i)})_{j,k}}(\boldsymbol{\theta}) && (3.178)\\
&= \sum_a [\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})]_a \frac{\partial (\vec{z}^{(i)})_a}{\partial (W^{(i)})_{j,k}} && (3.179)\\
&= \sum_a [\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})]_a \cdot \begin{cases} [\vec{\sigma}^{(i)}(\vec{z}^{(i-1)})]_k, & \text{if } a = j \text{ and } i \in \{1, \dots, m\} \\ x_k, & \text{if } a = j \text{ and } i = 0 \\ 0, & \text{otherwise} \end{cases} && (3.180)\\
&= \begin{cases} [\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})]_j [\vec{\sigma}^{(i)}(\vec{z}^{(i-1)})]_k, & \text{if } i \in \{1, \dots, m\} \\ [\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})]_j [\vec{x}]_k, & \text{if } i = 0 \end{cases} && (3.181)\\
&= \begin{cases} [\{\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})\} \vec{\sigma}^{(i)}(\vec{z}^{(i-1)})^\top]_{j,k}, & \text{if } i \in \{1, \dots, m\} \\ \{\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})\} \vec{x}^\top]_{j,k}, & \text{if } i = 0. \end{cases} && (3.182)
\end{aligned}$$

This gives

$$\nabla_{W^{(i)}} L(\boldsymbol{\theta}) = \begin{cases} [\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})] \vec{\sigma}^{(i)}(\vec{z}^{(i-1)})^\top, & \text{if } i \in \{1, \dots, m\} \\ [\nabla_{\vec{z}^{(i)}} L(\boldsymbol{\theta})] \vec{x}^\top, & \text{if } i = 0. \end{cases} \tag{3.183}$$

In combination with the expression for $\nabla_{\vec{b}^{(i)}}$ from Example 62, we can efficiently compute $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, and are able to train our neural network via gradient-based optimization methods such as gradient descent.

Chapter 4

Linear and Ridge Regression

Relevant sections of the textbooks:

- [2] Chapter 6.

4.1 Impact of Perturbations on Linear Regression

Before we start thinking about generic convex analysis, we will first study a particularly instructive, useful, and interesting linear-algebraic convex optimization problem.

Let $A \in \mathbb{R}^{n \times n}$ be invertible and $\vec{y} \in \mathbb{R}^n$. Consider the generic linear system $A\vec{x} = \vec{y}$, perhaps representing measurements of some physical system. There is exactly one $\vec{x}$ which solves this system — that being $\vec{x} = A^{-1}\vec{y}$. We want to understand how sensitive this system is to perturbations in the output. That is, if $\vec{y}$ is perturbed by $\vec{\delta}_{\vec{y}}$ for $\left\|\vec{\delta}_{\vec{y}}\right\|_2$ small (say, representing noise in the measurements), then the $\vec{x}$ that solves the system is *also* perturbed, say by $\vec{\delta}_{\vec{x}}$. So in the end, we have

$$A(\vec{x} + \vec{\delta}_{\vec{x}}) = (\vec{y} + \vec{\delta}_{\vec{y}}). \tag{4.1}$$

We want to compute the relative change in $\vec{x}$, that is, $\frac{\left\|\vec{\delta}_{\vec{x}}\right\|_2}{\|\vec{x}\|_2}$, in terms of $\left\|\vec{\delta}_{\vec{y}}\right\|_2$, as well as other properties of the system. In the context of our physical measurement system, we would much rather have this ratio be *small*; this means that the solutions to the equations governing our physical system are *robust* to measurement errors, thus assuring us that our model is relatively accurate to the real-life physical system. Thus, at the least we want to upper-bound $\frac{\left\|\vec{\delta}_{\vec{x}}\right\|_2}{\|\vec{x}\|_2}$.

The first part of upper-bounding $\frac{\left\|\vec{\delta}_{\vec{x}}\right\|_2}{\|\vec{x}\|_2}$ is to upper-bound $\left\|\vec{\delta}_{\vec{x}}\right\|_2$. We have

$$\begin{aligned}
A(\vec{x} + \vec{\delta}_{\vec{x}}) &= \vec{y} + \vec{\delta}_{\vec{y}} && (4.2)\\
A\vec{x} + A\vec{\delta}_{\vec{x}} &= \vec{y} + \vec{\delta}_{\vec{y}} && (4.3)\\
A\vec{\delta}_{\vec{x}} &= \vec{\delta}_{\vec{y}} && (4.4)\\
\vec{\delta}_{\vec{x}} &= A^{-1}\vec{\delta}_{\vec{y}}. && (4.5)
\end{aligned}$$

Then by taking norms on both sides,

$$\begin{aligned}
\left\|\vec{\delta}_{\vec{x}}\right\|_2 &= \left\|A^{-1}\vec{\delta}_{\vec{y}}\right\|_2 && (4.6)\\
&\leq \max_{\substack{\vec{z} \in \mathbb{R}^n \\ \|\vec{z}\|_2 = \left\|\vec{\delta}_{\vec{y}}\right\|_2}} \left\|A^{-1}\vec{z}\right\|_2 && (4.7)
\end{aligned}$$

$$= \left(\max_{\substack{\vec{z}\in\mathbb{R}^n \\ \|\vec{z}\|_2=1}} \left\|A^{-1}\vec{z}\right\|_2\right)\left\|\vec{\delta}_{\vec{y}}\right\|_2 \tag{4.8}$$

$$= \left\|A^{-1}\right\|_2 \left\|\vec{\delta}_{\vec{y}}\right\|_2 . \tag{4.9}$$

In order to upper-bound $\frac{\left\|\vec{\delta}_{\vec{x}}\right\|_2}{\|\vec{x}\|_2}$, we also need to lower-bound $\|\vec{x}\|_2$. Applying the same matrix norm inequality to the regular linear system $A\vec{x} = \vec{y}$ gives

$$A\vec{x} = \vec{y} \tag{4.10}$$

$$\|A\vec{x}\|_2 = \|\vec{y}\|_2 \tag{4.11}$$

$$\|A\|_2 \|\vec{x}\|_2 \geq \|\vec{y}\|_2 \tag{4.12}$$

$$\|\vec{x}\|_2 \geq \frac{\|\vec{y}\|}{\|A\|_2} \tag{4.13}$$

where $\|A\|_2 \neq 0$ because A is invertible. Plugging in both bounds, we have

$$\frac{\left\|\vec{\delta}_{\vec{x}}\right\|_2}{\|\vec{x}\|_2} \leq \frac{\left\|A^{-1}\right\|_2 \left\|\vec{\delta}_{\vec{y}}\right\|_2}{\|\vec{y}\|_2 / \|A\|_2} \tag{4.14}$$

$$= \|A\|_2 \left\|A^{-1}\right\|_2 \cdot \frac{\left\|\vec{\delta}_{\vec{y}}\right\|_2}{\|\vec{y}\|_2}. \tag{4.15}$$

Thus we've bounded the relative change in $\vec{x}$ by the relative change in $\vec{y}$. If the relative change in $\vec{y}$ is small, then the relative change in $\vec{x}$ will be small, and so on. But we'd like to say something more about $\|A\|_2 \left\|A^{-1}\right\|_2$, and indeed we can:

$$\|A\|_2 \left\|A^{-1}\right\|_2 = \sigma_1\{A\} \cdot \sigma_1\{A^{-1}\} = \frac{\sigma_1\{A\}}{\sigma_n\{A\}}, \tag{4.16}$$

where again, $\sigma_n\{A\} \neq 0$ because A is invertible. This quantity

$$\kappa(A) \doteq \frac{\sigma_1\{A\}}{\sigma_n\{A\}} \tag{4.17}$$

is called the *condition number* of a matrix. In general, for non-invertible systems, this can be infinite, but has the same definition.

Definition 85 (Condition Number)

Let $A \in \mathbb{R}^{n\times n}$. The *condition number* of A, denoted $\kappa(A)$, is given by

$$\kappa(A) \doteq \frac{\sigma_1\{A\}}{\sigma_n\{A\}}. \tag{4.18}$$

If $\kappa(A)$ is large, then even a small change in our measurement $\vec{y}$ will result in a huge change in our variable $\vec{x}$. If $\kappa(A)$ is small, then large changes in our measurement $\vec{y}$ result in small changes to our variable $\vec{x}$.

It seems unlikely that in general, the equations that define our system will be square. Most likely we will have a least-squares type *tall* system. But this is resolved by using the so-called *normal equations* to represent the least squares solution:

$$A^{\top}A\vec{x} = A^{\top}\vec{y}. \tag{4.19}$$

The condition number of this linear system is $\kappa(A^\top A)$. Since $A^\top A$ is symmetric and positive semidefinite, its eigenvalues are also its singular values, and so we have

$$\kappa(A^\top A) = \frac{\lambda_{\max}\{A^\top A\}}{\lambda_{\min}\{A^\top A\}}. \tag{4.20}$$

4.2 Ridge Regression

Sometimes we have least squares systems with $\kappa(A^\top A) = \infty$, or even finite but very large. How do we make this system solvable, robust, or even better conditioned? The answer goes like the following: suppose we could add some number λ to all eigenvalues of $A^\top A$. Then, since $\lambda_1\{A^\top A\}$ and $\lambda_n\{A^\top A\}$ go up by the same amount, $\lambda_n\{A^\top A\}$ becomes a larger fraction of $\lambda_1\{A^\top A\}$, so the condition number $\kappa(A^\top A)$ becomes lower.

This is perhaps easier to see with a numerical example, which we provide now. Suppose that $A^\top A$ has $\lambda_1\{A^\top A\} = 5$ and $\lambda_n\{A^\top A\} = 0.01$. Then $\kappa(A^\top A) = 500$. But if we add 3 to all eigenvalues of $A^\top A$, then $\lambda_1\{A^\top A\} = 8$ and $\lambda_n\{A^\top A\} = 3.01$, so $\kappa(A^\top A) = \frac{8}{3.01} \approx 2.65$. This is a much better-conditioned problem.

The question is now how to add λ to all eigenvalues of $A^\top A$. Using the shift property of eigenvalues, we see that we can add λI to $A^\top A$, so that instead of solving the system $A^\top A\vec{x} = A^\top\vec{y}$ we instead solve the system $(A^\top A + \lambda I)\vec{x} = A^\top\vec{y}$. The problem of finding $\vec{x}$ which solves this system is called *ridge regression*. It turns out to be equivalent to the following formulation.

Theorem 86 (Ridge Regression)

Let $A \in \mathbb{R}^{m\times n}$, $\vec{y} \in \mathbb{R}^m$, and $\lambda > 0$. The unique solution to the *ridge regression* problem

$$\min_{\vec{x}\in\mathbb{R}^n} \left\{\|A\vec{x} - \vec{y}\|_2^2 + \lambda\|\vec{x}\|_2^2\right\} \tag{4.21}$$

is given by

$$\vec{x}^\star = (A^\top A + \lambda I)^{-1}A^\top\vec{y}. \tag{4.22}$$

Proof. Let $f(\vec{x}) \doteq \|A\vec{x} - \vec{y}\|_2^2 + \lambda\|\vec{x}\|_2^2$. By taking gradients, we get

$$\begin{aligned}
\nabla_{\vec{x}} f(\vec{x}) &= \nabla_{\vec{x}}\left\{\|A\vec{x} - \vec{y}\|_2^2 + \lambda\|\vec{x}\|_2^2\right\} && (4.23)\\
&= \nabla_{\vec{x}}\{\vec{x}^\top A^\top A\vec{x} - 2\vec{y}^\top A\vec{x} + \vec{y}^\top\vec{y} + \lambda\vec{x}^\top\vec{x}\} && (4.24)\\
&= 2A^\top A\vec{x} - 2A^\top\vec{y} + 2\lambda\vec{x} && (4.25)\\
&= 2(A^\top A + \lambda I)\vec{x} - 2A^\top\vec{y}. && (4.26)
\end{aligned}$$

Thus we get that the optimal point is determined by solving the linear system

$$(A^\top A + \lambda I)\vec{x} = A^\top\vec{y}. \tag{4.27}$$

Since $A^\top A$ is PSD and $\lambda > 0$, we have $A^\top A + \lambda I$ is PD and thus invertible. Therefore

$$\vec{x}^\star = (A^\top A + \lambda I)^{-1}A^\top\vec{y} \tag{4.28}$$

is the unique solution to the above linear system and therefore the unique solution to the optimization problem. □

Note that we haven't proved that a convex function (such as the above ridge regression objective) is minimized when its derivative is 0; we prove this in subsequent lectures, but for now let us take it for granted.

Proof. Another way to solve the same problem is to consider the augmented system

$$\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{y} \\ \vec{0} \end{bmatrix}. \tag{4.29}$$

This augmented matrix has full column rank, so we can use the least squares solution to get a unique solution for $\vec{x}$. We get

$$\begin{aligned}
\vec{x} &= \left(\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}^{\top} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}\right)^{-1} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}^{\top} \begin{bmatrix} \vec{y} \\ \vec{0} \end{bmatrix} && (4.30) \\
&= \left(\begin{bmatrix} A^{\top} & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}\right)^{-1} \begin{bmatrix} A^{\top} & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} \vec{y} \\ \vec{0} \end{bmatrix} && (4.31) \\
&= \left(\begin{bmatrix} A^{\top} & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}\right)^{-1} \left(A^{\top} \vec{y} + \sqrt{\lambda} I \cdot \vec{0}\right) && (4.32) \\
&= \left(\begin{bmatrix} A^{\top} & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}\right)^{-1} A^{\top} \vec{y} && (4.33) \\
&= \left(A^{\top} A + \lambda I\right)^{-1} A^{\top} \vec{y}. && (4.34)
\end{aligned}$$

□

In the ridge regression objective

$$\left\|A\vec{x} - \vec{b}\right\|_2^2 + \lambda \left\|\vec{x}\right\|_2^2, \tag{4.35}$$

the second term $\lambda \left\|\vec{x}\right\|_2^2$ is called a *regularizer*; this is because it *regulates* or *regularizes* our problem by making it better-conditioned.

4.3 Principal Components Regression

We can gain more understanding of the ridge regression solution by looking at it through the SVD of A. Indeed, let $A = U\Sigma V^{\top}$. Then the ridge regression solution is

$$\begin{aligned}
x^{\star} &= \left(A^{\top} A + \lambda I\right)^{-1} A^{\top} \vec{y} && (4.36) \\
&= \left((U\Sigma V^{\top})^{\top}(U\Sigma V^{\top}) + \lambda I\right)^{-1} (U\Sigma V^{\top})^{\top} \vec{y} && (4.37) \\
&= \left(V\Sigma^{\top} U^{\top} U\Sigma V^{\top} + \lambda I\right)^{-1} V\Sigma^{\top} U^{\top} \vec{y} && (4.38) \\
&= \left(V\Sigma^{\top}\Sigma V^{\top} + \lambda I\right)^{-1} V\Sigma^{\top} U^{\top} \vec{y} && (4.39) \\
&= \left(V\Sigma^{\top}\Sigma V^{\top} + V(\lambda I)V^{\top}\right)^{-1} V\Sigma^{\top} U^{\top} \vec{y} && (4.40) \\
&= \left(V\left(\Sigma^{\top}\Sigma + \lambda I\right)V^{\top}\right)^{-1} V\Sigma^{\top} U^{\top} \vec{y} && (4.41) \\
&= V\left(\Sigma^{\top}\Sigma + \lambda I\right)^{-1} V^{\top} V\Sigma^{\top} U^{\top} \vec{y} && (4.42) \\
&= V\left(\Sigma^{\top}\Sigma + \lambda I\right)^{-1} \Sigma^{\top} U^{\top} \vec{y} && (4.43) \\
&= V \begin{bmatrix} (\Sigma_r^2 + \lambda I)^{-1}\Sigma_r & 0 \\ 0 & 0 \end{bmatrix} U^{\top} \vec{y}. && (4.44)
\end{aligned}$$

Looking at the middle matrix a bit more, we see that

$$(\Sigma_r^2 + \lambda I)^{-1}\Sigma_r = \begin{bmatrix} \frac{\sigma_1\{A\}}{\sigma_1\{A\}^2+\lambda} & & \\ & \ddots & \\ & & \frac{\sigma_n\{A\}}{\sigma_n\{A\}^2+\lambda} \end{bmatrix}.$$

Thus, we get

$$\begin{aligned} \vec{x}^\star &= V \begin{bmatrix} (\Sigma_r^2 + \lambda I)^{-1}\Sigma_r & 0 \\ 0 & 0 \end{bmatrix} U^\top \vec{y} && (4.45) \\ &= \left(\sum_{i=1}^{r} \frac{\sigma_i\{A\}}{\sigma_i\{A\}^2 + \lambda} \vec{v}_i \vec{u}_i^\top \right) \vec{y} && (4.46) \\ &= \sum_{i=1}^{r} \frac{\sigma_i\{A\}}{\sigma_i\{A\}^2 + \lambda} (\vec{u}_i^\top \vec{y}) \cdot \vec{v}_i. && (4.47) \end{aligned}$$

To understand what λ is doing here, we contrast two examples. Let $A \in \mathbb{R}^{n \times 3}$ for some large $n \gg 3$.

Suppose first that $\sigma_1\{A\} = \sigma_2\{A\} = \sigma_3\{A\} = 1$. Then

$$\begin{aligned} \vec{x}^\star &= \frac{\sigma_1\{A\}}{\sigma_1\{A\}^2 + \lambda}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{\sigma_2\{A\}}{\sigma_2\{A\}^2 + \lambda}(\vec{u}_2^\top \vec{y})\vec{v}_2 + \frac{\sigma_3\{A\}}{\sigma_3\{A\}^2 + \lambda}(\vec{u}_3^\top \vec{y})\vec{v}_3 && (4.48) \\ &= \frac{1}{1+\lambda}\{(\vec{u}_1^\top \vec{y})\vec{v}_1 + (\vec{u}_2^\top \vec{y})\vec{v}_2 + (\vec{u}_3^\top \vec{y})\vec{v}_3\} && (4.49) \\ &= \frac{1}{1+\lambda}\vec{\tilde{x}} && (4.50) \end{aligned}$$

where $\vec{\tilde{x}}$ is the solution of the corresponding least squares linear regression problem, namely the ridge problem with $\lambda = 0$. In this way, the λ parameter decays the solution in each principal direction equally, pulling the whole $\vec{\tilde{x}}$ vector towards 0. This is interesting precisely because a first-level examination of the ridge regression objective function — and namely the $\|\vec{x}\|_2^2$ term, which by itself penalizes every direction of $\vec{x}$ equally — may make it seem like this is always the case, but it turns out to not be, as we will see shortly.

Now suppose that $\sigma_1\{A\} = 100$, $\sigma_2\{A\} = 10$, and $\sigma_3\{A\} = 1$. Then

$$\begin{aligned} \vec{x}^\star &= \frac{\sigma_1\{A\}}{\sigma_1\{A\}^2 + \lambda}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{\sigma_2\{A\}}{\sigma_2\{A\}^2 + \lambda}(\vec{u}_2^\top \vec{y})\vec{v}_2 + \frac{\sigma_3\{A\}}{\sigma_3\{A\}^2 + \lambda}(\vec{u}_3^\top \vec{y})\vec{v}_3 && (4.51) \\ &= \frac{100}{10000+\lambda}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{10}{100+\lambda}(\vec{u}_2^\top \vec{y})\vec{v}_2 + \frac{1}{1+\lambda}(\vec{u}_3^\top \vec{y})\vec{v}_3 && (4.52) \\ &= \frac{1}{100+\lambda/100}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{1}{10+\lambda/10}(\vec{u}_2^\top \vec{y})\vec{v}_2 + \frac{1}{1+\lambda}(\vec{u}_3^\top \vec{y})\vec{v}_3. && (4.53) \end{aligned}$$

Thus, the different terms are now impacted differently based on λ; in particular, to impact the first term by a certain amount, one needs to change λ by 100 times the amount required to change the last term. Namely, if we set λ to be large, say $\lambda = 10000$, the coefficient of the first term becomes $1/110$, while the coefficient of the last term becomes $1/10001$ which is much lower. More generally, for a larger example, setting λ to be large effectively zeros out the last few terms while effectively not changing the first few terms. Thus, setting λ to be large effectively performs a "soft thresholding" of the singular values, making the terms associated with smaller singular values be nearly 0 while preserving the terms associated with larger singular values. More quantitatively, for large λ, we have

$$\vec{x}^\star = \frac{1}{100+\lambda/100}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{1}{10+\lambda/10}(\vec{u}_2^\top \vec{y})\vec{v}_2 + \frac{1}{1+\lambda}(\vec{u}_3^\top \vec{y})\vec{v}_3 \qquad (4.54)$$

$$\approx \frac{1}{100+\lambda/100}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{1}{10+\lambda/10}(\vec{u}_2^\top \vec{y})\vec{v}_2 \tag{4.55}$$

and for even larger λ we simply have

$$\vec{x}^\star = \frac{1}{100+\lambda/100}(\vec{u}_1^\top \vec{y})\vec{v}_1 + \frac{1}{10+\lambda/10}(\vec{u}_2^\top \vec{y})\vec{v}_2 + \frac{1}{1+\lambda}(\vec{u}_3^\top \vec{y})\vec{v}_3 \tag{4.56}$$

$$\approx \frac{1}{100+\lambda/100}(\vec{u}_1^\top \vec{y})\vec{v}_1. \tag{4.57}$$

Since the terms form a linear combination of the $\vec{v}_i$, setting such terms associated with small singular values to (nearly) 0 is similar to performing PCA, where we only use the $\vec{v}_i$ associated with the largest few singular values. Thus our conclusion is — ridge regression behaves qualitatively similar to a soft form of PCA.

4.4 Tikhonov Regression

Recall that our earlier augmented system

$$\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{y} \\ \vec{0} \end{bmatrix} \tag{4.58}$$

which had full column rank, tried to find a $\vec{x}$ such that $A\vec{x} \approx \vec{y}$ while $\vec{x} \approx \vec{0}$ — in other words, $\vec{x}$ that is small. Suppose that we wanted to instead enforce that $\vec{x}$ were close to some other vector $\vec{x}_0 \in \mathbb{R}^n$. Then we would set up the system

$$\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{y} \\ \vec{x}_0 \end{bmatrix}. \tag{4.59}$$

This would yield the least-squares type objective function

$$\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x} - \vec{x}_0\|_2^2. \tag{4.60}$$

The final generalization of this is to put different weights on each row of $A\vec{x} - \vec{y}$ and $\vec{x} - \vec{x}_0$. If, for example, we really want to get row i of $A\vec{x}$ close to b_i, we can multiply the squared difference $(A\vec{x} - \vec{y})_i^2$ by a large weight in the loss function, and the solutions will bias towards ensuring that $(A\vec{x} - \vec{y})_i \approx 0$. Similarly, if we really are sure that the true $\vec{x}$ has i^{th} coordinate $(\vec{x}_0)_i$, then we can attach a large weight to the difference $(\vec{x} - \vec{x}_0)_i^2$ as well. Mathematically, this gives us the following objective function:

$$\|W_1(A\vec{x} - \vec{y})\|_2^2 + \|W_2(\vec{x} - \vec{x}_0)\|_2^2, \tag{4.61}$$

where $W_1 \in \mathbb{R}^{m \times m}$ and $W_2 \in \mathbb{R}^{n \times n}$ are diagonal matrices representing the weights. Notice how this is a generalization of ridge regression with $W_1 = I$, $W_2 = \sqrt{\lambda} I$, and $\vec{x}_0 = \vec{0}$. This general regression is called Tikhonov regression.

Theorem 87 (Tikhonov Regression)

Let $A \in \mathbb{R}^{m \times n}$, $\vec{x}_0 \in \mathbb{R}^n$, and $\vec{y} \in \mathbb{R}^m$, and let $W_1 \in \mathbb{R}^{m \times m}$ and $W_2 \in \mathbb{R}^{n \times n}$ be diagonal. Then the unique solution to the *Tikhonov regression* problem

$$\min_{\vec{x} \in \mathbb{R}^n} \left\{ \|W_1(A\vec{x} - \vec{y})\|_2^2 + \|W_2(\vec{x} - \vec{x}_0)\|_2^2 \right\} \tag{4.62}$$

is given by

$$\vec{x}^\star = (A^\top W_1^2 A + W_2^2)^{-1}(A^\top W_1^2 \vec{y} + W_2^2 \vec{x}_0). \tag{4.63}$$

Proof. Left as exercise. □

This expression looks complicated, so we do a sanity-check; if $W_1 = I$, $W_2 = \sqrt{\lambda}I$, and $\vec{x}_0 = \vec{0}$, then we get exactly the ridge regression solution.

4.5 Maximum Likelihood Estimation (MLE)

Previously, we talked about incorporating side information (like $\vec{x} = \vec{x}_0$) deterministically. Now we discuss a way to incorporate probabilistic information into our model.

Namely, suppose that the rows of our A matrix are vectors $\vec{a}_1, \dots, \vec{a}_m \in \mathbb{R}^n$, and that the entries of our $\vec{y}$ vector are $y_1, \dots, y_m \in \mathbb{R}$. Now suppose we have the probabilistic model

$$y_i = \vec{a}_i^\top \vec{x} + w_i, \qquad \forall i \in \{1, \dots, m\} \tag{4.64}$$

where $w_1, \dots, w_n$ are independent Gaussian random variables; in particular, $w_i \sim \mathcal{N}(0, \sigma_i^2)$. Or in short, we have

$$\vec{y} = A\vec{x} + \vec{w} \tag{4.65}$$

where $\vec{w} = \begin{bmatrix} w_1 & \cdots & w_m \end{bmatrix}^\top \in \mathbb{R}^m$. In this case, we say that $\vec{w} \sim \mathcal{N}(\vec{0}, \Sigma_{\vec{w}})$ where $\Sigma_{\vec{w}} \doteq \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_m^2\right)$

In this setup, the *maximum likelihood estimate* (MLE) for $\vec{x}$ turns out to be exactly a solution to a Tikhonov regression problem. The maximum likelihood estimate is the parameter choice which makes the data most likely, in that it has the highest probability or probability density out of all choices of the parameter. It is a meaningful and popular statistical estimator; thus the fact that we can reduce its computation to a ridge regression-type problem is both interesting and useful.

Henceforth, we use p to denote probability densities, and use $p_{\vec{x}}$ to denote probability densities for a fixed value of $\vec{x}$. In the above model, $\vec{x}$ is not a random variable, so it doesn't quite make formal sense to condition on it (though — spoilers! — we will soon put a probabilistic prior on it, and then it makes sense to condition).

Proposition 88 (MLE as Tikhonov Regression)

In the above probabilistic model, we have

$$\operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} p_{\vec{x}}(\vec{y}) = \operatorname*{argmin}_{\vec{x} \in \mathbb{R}^n} \left\| \Sigma_{\vec{w}}^{-1/2}(A\vec{x} - \vec{y}) \right\|_2^2. \tag{4.66}$$

Proof. Since the logarithm is monotonically increasing, $\operatorname{argmax}_{\vec{x}} f(\vec{x}) = \operatorname{argmax}_{\vec{x}} \log(f(\vec{x}))$ for all functions f, and so

$$\begin{aligned}
\operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} p_{\vec{x}}(\vec{y}) &= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \log(p_{\vec{x}}(\vec{y})) && (4.67)\\
&= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \log\left(\prod_{i=1}^m p_{\vec{x}}(y_i)\right) && (4.68)\\
&= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^m \log(p_{\vec{x}}(y_i)) && (4.69)\\
&= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^m \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2}\right)\right) && (4.70)
\end{aligned}$$

$$= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \left\{ \underbrace{\log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right)}_{\text{independent of } \vec{x}} + \log\left(\exp\left(-\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2}\right)\right) \right\} \tag{4.71}$$

$$= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \left\{ \log\left(\exp\left(-\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2}\right)\right) \right\} \tag{4.72}$$

$$= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \left\{ -\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2} \right\} \tag{4.73}$$

$$= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \left\{ -\frac{1}{2} \sum_{i=1}^{m} \frac{(y_i - \vec{a}_i^\top \vec{x})^2}{\sigma_i^2} \right\} \tag{4.74}$$

$$= \operatorname*{argmin}_{\vec{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \frac{(y_i - \vec{a}_i^\top \vec{x})^2}{\sigma_i^2} \tag{4.75}$$

$$= \operatorname*{argmin}_{\vec{x} \in \mathbb{R}^n} \left\| \Sigma_{\vec{w}}^{-1/2} (A\vec{x} - \vec{y}) \right\|_2^2 . \tag{4.76}$$

□

4.6 Maximum A Posteriori Estimation (MAP)

Now we consider the same probabilistic model as above, except this time suppose that we believe $\vec{x}$ is also random, in the sense that

$$x_j = \mu_j + v_j, \qquad \forall j \in \{1, \ldots, n\} \tag{4.77}$$

where $v_1, \ldots, v_n$ are independent Gaussian random variables; in particular $v_j \sim \mathcal{N}(0, \tau_j^2)$. Or in short, we have

$$\vec{x} = \vec{x}_0 + \vec{v} \tag{4.78}$$

where $\vec{v} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^\top \in \mathbb{R}^n$ is distributed as $\vec{v} \sim \mathcal{N}(\vec{0}, \Sigma_{\vec{v}})$, where $\Sigma_{\vec{v}} \doteq \operatorname{diag}(\tau_1^2, \ldots, \tau_n^2)$.

In this setup, the maximum likelihood estimate may still be useful, but another quantity that is perhaps more relevant is the *maximum a posteriori estimate* (MAP). The MAP estimate is the value of $\vec{x}$ which is most likely, i.e., having the highest conditional probability or conditional probability density, conditioned on the observed data. It is also a meaningful and popular statistical estimator. It turns out that we can derive a similar result as in the MLE case.

Theorem 89 (MAP as Tikhonov Regression)
In the above probabilistic model, we have

$$\operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} p(\vec{x} \mid \vec{y}) = \operatorname*{argmin}_{\vec{x} \in \mathbb{R}^n} \left\{ \left\| \Sigma_{\vec{w}}^{-1/2} (A\vec{x} - \vec{y}) \right\|_2^2 + \left\| \Sigma_{\vec{v}}^{-1/2} (\vec{x} - \vec{x}_0) \right\|_2^2 \right\} . \tag{4.79}$$

Proof. Using Bayes' rule and the computations from before, we have

$$\operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} p(\vec{x} \mid \vec{y}) \tag{4.80}$$

$$= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \log(p(\vec{x} \mid \vec{y})) \tag{4.81}$$

$$= \operatorname*{argmax}_{\vec{x} \in \mathbb{R}^n} \log\left(\frac{p(\vec{y} \mid \vec{x})p(\vec{x})}{p(\vec{y})}\right) \tag{4.82}$$

$$= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \left\{ \log(p(\vec{y} \mid \vec{x})) + \log(p(\vec{x})) - \underbrace{\log(p(\vec{y}))}_{\text{independent of } \vec{x}} \right\} \tag{4.83}$$

$$= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \left\{ \log(p(\vec{y} \mid \vec{x})) + \log(p(\vec{x})) \right\} \tag{4.84}$$

$$= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \log\left(\left\{ \prod_{i=1}^{m} p(y_i \mid \vec{x}) \right\} \cdot \left\{ \prod_{j=1}^{n} p(x_j) \right\} \right) \tag{4.85}$$

$$= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{m} \log(p(y_i \mid \vec{x})) + \sum_{j=1}^{n} \log(p(x_j)) \right\} \tag{4.86}$$

$$= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2} \right) \right) + \sum_{j=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{(x_j - (\vec{x}_0)_j)^2}{2\tau_j^2} \right) \right) \right\} \tag{4.87}$$

$$= \operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{m} \left(-\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{2\sigma_i^2} \right) + \sum_{j=1}^{n} \left(-\frac{(x_j - (\vec{x}_0)_j)^2}{2\tau_j^2} \right) \right\} \tag{4.88}$$

$$= \operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{m} \left(\frac{(y_i - \vec{a}_i^\top \vec{x})^2}{\sigma_i^2} \right) + \sum_{j=1}^{n} \left(\frac{(x_j - (\vec{x}_0)_j)^2}{\tau_j^2} \right) \right\} \tag{4.89}$$

$$= \operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \left\{ \left\| \Sigma_{\vec{w}}^{-1/2} (A\vec{x} - \vec{y}) \right\|_2^2 + \left\| \Sigma_{\vec{v}}^{-1/2} (\vec{x} - \vec{x}_0) \right\|_2^2 \right\} \tag{4.90}$$

as desired. □

Chapter 5

Convexity

Relevant sections of the textbooks:

- [1] Chapter 4.
- [2] Chapter 8.

5.1 Convex Sets

5.1.1 Basics

First, we want to define a special type of linear combination called a *convex combination*.

Definition 90 (Convex Combination)
Let $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$. The sum

$$\vec{x} = \sum_{i=1}^{k} \theta_i \vec{x}_i \tag{5.1}$$

is a *convex combination* of $\vec{x}_1, \ldots, \vec{x}_k$ if each $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$.

We can think of each θ_i as a weight on the corresponding $\vec{x}_i$. Since they are non-negative numbers which sum to 1, we can also interpret them as probabilities.

Definition 91 (Convex Set)
Let $C \subseteq \mathbb{R}^n$. We say that C is a *convex set* if it is closed under convex combinations: for all $\vec{x}_1, \vec{x}_2 \in C$ and all $\theta \in [0, 1]$, we have $\theta\vec{x}_1 + (1 - \theta)\vec{x}_2 \in C$.

Geometrically, a set C is convex if for every two points $\vec{x}_1, \vec{x}_2 \in C$, the line segment $\{\theta\vec{x}_1 + (1 - \theta)\vec{x}_2 \mid \theta \in [0, 1]\}$ is contained in C. This means that, for example, the midpoint between $\vec{x}_1$ and $\vec{x}_2$, i.e., $\frac{1}{2}\vec{x}_1 + \frac{1}{2}\vec{x}_2$, is contained in C, as well as the point $\frac{1}{3}$ of the way from $\vec{x}_1$ to $\vec{x}_2$, i.e., $\frac{2}{3}\vec{x}_1 + \frac{1}{3}\vec{x}_2$, etc. More generally, as we vary θ, we go along the line segment connecting $\vec{x}_1$ and $\vec{x}_2$.