Summary

Key topics.

▶ Familiarity with form of basic network gradient.
▶ Deep network initialization.
▶ Minibatches.
▶ Momentum.

Next time: convexity.
Part 2: convexity
Deep networks are not convex in their parameters.

*Why study convexity?*

- Convexity is pervasive in ML and mathematics; e.g., our losses for deep learning are still convex.
- Convexity exemplifies nice “local-to-global” structure.
6. Convex sets and functions
Convex sets

A set $S$ is **convex** if, for every pair of points $\{x, x'\}$ in $S$, the line segment between $x$ and $x'$ is also contained in $S$. ($\{x, x'\} \in S \implies [x, x'] \in S$.)
Convex sets

A set $S$ is *convex* if, for every pair of points \( \{x, x'\} \) in $S$, the line segment between $x$ and $x'$ is also contained in $S$. (\( \{x, x'\} \in S \implies [x, x'] \in S \).)

![Convex sets](image)

**Examples:**

- All of $\mathbb{R}^d$.
- Empty set.
- Half-spaces: \( \{x \in \mathbb{R}^d : \mathbf{a}^T x \leq b\} \).
- Intersections of convex sets.
- Polyhedra: \( \left\{ x \in \mathbb{R}^d : \mathbf{A} x \leq \mathbf{b} \right\} = \bigcap_{i=1}^{m} \left\{ x \in \mathbb{R}^d : \mathbf{a}_i^T x \leq b_i \right\} \).
- Convex hulls:
  \[
  \text{conv}(S) := \left\{ \sum_{i=1}^{k} \alpha_i x_i : k \in \mathbb{N}, \ x_i \in S, \ \alpha_i \geq 0, \ \sum_{i=1}^{k} \alpha_i = 1 \right\}. 
  \]
  (Infinite convex hulls: intersection of all convex supersets.)
Convex functions from convex sets

The *epigraph* of a function $f$ is the area above the curve:

$$\text{epi}(f) := \left\{ (x, y) \in \mathbb{R}^{d+1} : y \geq f(x) \right\}.$$

A function is convex if its epigraph is convex.
Convex functions (standard definition)

A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is **convex** if for any $x, x' \in \mathbb{R}^d$ and $\alpha \in [0, 1],$

$$f ((1 - \alpha)x + \alpha x') \leq (1 - \alpha) \cdot f(x) + \alpha \cdot f(x').$$

**Examples:**

- $f(x) = cx$ for any $c > 0$ (on $\mathbb{R}$)
- $f(x) = |x| c$ for any $c \geq 1$ (on $\mathbb{R}$)
- $f(x) = b^T x$ for any $b \in \mathbb{R}^d$.
- $f(x) = \|x\|$ for any norm $\|\cdot\|$.  
- $f(x) = x^T A x$, which approximates $\max_i x_i$. 

$f$ is not convex

$f$ is convex
Convex functions (standard definition)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for any $x, x' \in \mathbb{R}^d$ and $\alpha \in [0, 1]$,

$$f \left((1 - \alpha)x + \alpha x'\right) \leq (1 - \alpha) \cdot f(x) + \alpha \cdot f(x') .$$

**Examples:**

- $f(x) = c^x$ for any $c > 0$ (on $\mathbb{R}$)
- $f(x) = |x|^c$ for any $c \geq 1$ (on $\mathbb{R}$)
- $f(x) = b^T x$ for any $b \in \mathbb{R}^d$.
- $f(x) = \|x\|$ for any norm $\|\cdot\|$.
- $f(x) = x^T A x$ for symmetric positive semidefinite $A$.
- $f(x) = \ln \left(\sum_{i=1}^d \exp(x_i)\right)$, which approximates $\max_i x_i$. 
Example verification: norms

Is $f(x) = \|x\|$ convex?
Is \( f(x) = \|x\| \) convex?

Pick any \( \alpha \in [0, 1] \) and any \( x, x' \in \mathbb{R}^d \).
Is $f(x) = \|x\|$ convex?

Pick any $\alpha \in [0, 1]$ and any $x, x' \in \mathbb{R}^d$.

$$f ((1 - \alpha)x + \alpha x') = \|(1 - \alpha)x + \alpha x'\|$$
Example verification: norms

Is \( f(x) = \|x\| \) convex?

Pick any \( \alpha \in [0, 1] \) and any \( x, x' \in \mathbb{R}^d \).

\[
\begin{align*}
f((1 - \alpha)x + \alpha x') &= \|(1 - \alpha)x + \alpha x'\| \\
&\leq \|(1 - \alpha)x\| + \|\alpha x'\| \quad \text{(triangle inequality)}
\end{align*}
\]
Is $f(x) = \|x\|$ convex?

Pick any $\alpha \in [0, 1]$ and any $x, x' \in \mathbb{R}^d$.

\[
\begin{align*}
f ((1 - \alpha)x + \alpha x') &= \|(1 - \alpha)x + \alpha x'\| \\
&\leq \|(1 - \alpha)x\| + \|\alpha x'\| \quad \text{(triangle inequality)} \\
&= (1 - \alpha)\|x\| + \alpha \|x'\| \quad \text{(homogeneity)}
\end{align*}
\]
Example verification: norms

Is \( f(x) = \|x\| \) convex?

Pick any \( \alpha \in [0, 1] \) and any \( x, x' \in \mathbb{R}^d \).

\[
\begin{align*}
f((1 - \alpha)x + \alpha x') &= \|(1 - \alpha)x + \alpha x'\| \\
&\leq \|(1 - \alpha)x\| + \|\alpha x'\| \quad \text{(triangle inequality)} \\
&= (1 - \alpha)\|x\| + \alpha \|x'\| \quad \text{(homogeneity)} \\
&= (1 - \alpha)f(x) + \alpha f(x').
\end{align*}
\]

Yes, \( f \) is convex.
Operations preserving convexity

**Summations:** if \((f_1, \ldots, f_k)\) convex and \((\alpha_1, \ldots, \alpha_k)\) nonnegative,

\[
x \mapsto \alpha_1 f_1(x) + \cdots + \alpha_k f_k(x)
\]

is convex.

**Affine composition:** if \(f\) is convex, the for any \(A \in \mathbb{R}^{m \times d}\) and \(b \in \mathbb{R}^m\),

\[
x \mapsto f(Ax + b)
\]

is convex.

**Maxima:** if \((f_1, \ldots, f_k)\) are convex,

\[
x \mapsto \max_i f_i(x)
\]

is convex.

(Infinitely many functions: use a supremum.)
If $\ell$ is convex and the predictor is linear, then the empirical risk is convex:

- Define $\ell_i(w) = \ell(w^Tx_iy_i)$, convex since composition of convex and affine;
- thus the empirical risk

$$\hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^T x_i y_i) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(w)$$

is the nonnegative combination of convex functions, and convex.
7. Various forms of convexity
### Convexity of differentiable functions

#### Differentiable functions

If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then $f$ is convex if and only if

$$f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0)$$

for all $x, x_0 \in \mathbb{R}^d$.

**Note:** this implies *increasing slopes*:

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0.$$
Convexity of differentiable functions

Differentiable functions
If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then $f$ is convex if and only if

$$f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0)$$

for all $x, x_0 \in \mathbb{R}^d$.

Note: this implies increasing slopes:

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0.$$  

Twice-differentiable functions
If $f : \mathbb{R}^d \to \mathbb{R}$ is twice-differentiable, then $f$ is convex if and only if

$$\nabla^2 f(x) \succeq 0$$

for all $x \in \mathbb{R}^d$ (i.e., the Hessian, or matrix of second-derivatives, is positive semi-definite for all $x$).
Verifying convexity of differentiable functions

Is $f(x) = x^4$ convex?
Is \( f(x) = x^4 \) convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]
Verifying convexity of differentiable functions

Is \( f(x) = x^4 \) convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, \( f \) is convex.
Verifying convexity of differentiable functions

Is $f(x) = x^4$ convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, $f$ is convex.

Is $f(x) = e^{\langle a, x \rangle}$ convex?
Verifying convexity of differentiable functions

Is \( f(x) = x^4 \) convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, \( f \) is convex.

Is \( f(x) = e^{\langle \alpha, x \rangle} \) convex? Use first-order condition for convexity.

\[
\nabla f(x) = e^{\langle \alpha, x \rangle} \nabla \{ \langle \alpha, x \rangle \} = e^{\langle \alpha, x \rangle} \alpha \quad (\text{chain rule}).
\]
Verifying convexity of differentiable functions

Is $f(x) = x^4$ convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, $f$ is convex.

Is $f(x) = e^{\langle a, x \rangle}$ convex? Use first-order condition for convexity.

\[
\nabla f(x) = e^{\langle a, x \rangle} \nabla \{ \langle a, x \rangle \} = e^{\langle a, x \rangle} a \quad \text{(chain rule)}.
\]

Difference between $f$ and its affine approximation:

\[
f(x) - (f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle) = e^{\langle a, x \rangle} - \left( e^{\langle a, x_0 \rangle} + e^{\langle a, x_0 \rangle} \langle a, x - x_0 \rangle \right)
\]
Is \( f(x) = x^4 \) convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, \( f \) is convex.

Is \( f(x) = e^{\langle a, x \rangle} \) convex? Use first-order condition for convexity.

\[
\nabla f(x) = e^{\langle a, x \rangle} \nabla \{ \langle a, x \rangle \} = e^{\langle a, x \rangle} a \quad \text{(chain rule)}.
\]

Difference between \( f \) and its affine approximation:

\[
f(x) - \left( f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \right) = e^{\langle a, x \rangle} - \left( e^{\langle a, x_0 \rangle} + e^{\langle a, x_0 \rangle} \langle a, x - x_0 \rangle \right)
\]
\[
= e^{\langle a, x_0 \rangle} \left( e^{\langle a, x - x_0 \rangle} - (1 + \langle a, x - x_0 \rangle) \right)
\]
Verifying convexity of differentiable functions

Is \( f(x) = x^4 \) convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, \( f \) is convex.

Is \( f(x) = e^{\langle a, x \rangle} \) convex? Use first-order condition for convexity.

\[
\nabla f(x) = e^{\langle a, x \rangle} \nabla \{\langle a, x \rangle\} = e^{\langle a, x \rangle} a \quad (\text{chain rule}).
\]

Difference between \( f \) and its affine approximation:

\[
f(x) - (f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle) = e^{\langle a, x \rangle} - \left( e^{\langle a, x_0 \rangle} + e^{\langle a, x_0 \rangle} \langle a, x - x_0 \rangle \right)
\]
\[
= e^{\langle a, x_0 \rangle} \left( e^{\langle a, x-x_0 \rangle} - (1 + \langle a, x - x_0 \rangle) \right)
\]
\[
\geq 0 \quad (\text{because } 1 + z \leq e^z \text{ for all } z \in \mathbb{R}).
\]
Verifying convexity of differentiable functions

Is \( f(x) = x^4 \) convex? Use second-order condition for convexity.

\[
\frac{\partial}{\partial x} f(x) = 4x^3 \\
\frac{\partial^2}{\partial x^2} f(x) = 12x^2 \geq 0.
\]

Yes, \( f \) is convex.

Is \( f(x) = e^{\langle a, x \rangle} \) convex? Use first-order condition for convexity.

\[
\nabla f(x) = e^{\langle a, x \rangle} \nabla \{ \langle a, x \rangle \} = e^{\langle a, x \rangle} a \quad \text{(chain rule)}.
\]

Difference between \( f \) and its affine approximation:

\[
f(x) - \left( f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \right) = e^{\langle a, x \rangle} - \left( e^{\langle a, x_0 \rangle} + e^{\langle a, x_0 \rangle} \langle a, x - x_0 \rangle \right) \\
= e^{\langle a, x_0 \rangle} \left( e^{\langle a, x-x_0 \rangle} - (1 + \langle a, x - x_0 \rangle) \right) \\
\geq 0 \quad \text{(because } 1 + z \leq e^z \text{ for all } z \in \mathbb{R}).
\]

Yes, \( f \) is convex.
Strict convexity

Function values: \( \forall \mathbf{x}, \mathbf{y}, \forall \alpha \in [0, 1]: \)
\[
f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}).
\]

Derivatives: \( \forall \mathbf{x}, \mathbf{y}, \)
\[
f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).
\]

Hessians: \( \forall \mathbf{x}, \)
\[
\nabla^2 f(\mathbf{x}) \succeq 0.
\]
Strict convexity

Function values: \( \forall x \neq y, \forall \alpha \in (0, 1): \)

\[
f (\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y).
\]

Derivatives: \( \forall x \neq y, \)

\[
f(y) > f(x) + \nabla f(x)^\top (y - x).
\]

Hessians: \( \forall x, \)

\[
\nabla^2 f(x) \succ 0.
\]
\(\lambda\)-Strong-Convexity.

Function values: \(\forall x, y, \forall \alpha \in [0, 1]\)
\[
f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y).
\]

Derivatives: \(\forall x, y\)
\[
f(y) \geq f(x) + \nabla f(x)^\top (y - x).
\]

Hessians: \(\forall x,\)
\[
\nabla^2 f(x) \succeq 0.
\]
\(\lambda\)-Strong-Convexity.

Function values: \(\forall x, y, \forall \alpha \in [0, 1] \)

\[
f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\lambda\alpha(1 - \alpha)}{2} \|x - y\|^2.
\]

Derivatives: \(\forall x, y \)

\[
f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\lambda}{2} \|y - x\|^2.
\]

Hessians: \(\forall x, \)

\[
\nabla^2 f(x) \preceq \lambda I.
\]
Convexity of key losses.

Logistic loss \( z \mapsto \ln(1 + \exp(-z)) \) is **strictly convex**. (e.g., verify that second derivative is positive.)
Convexity of key losses.

**Logistic loss** \( z \mapsto \ln(1 + \exp(-z)) \) is **strictly convex**.
(e.g., verify that second derivative is positive.)

**Squared (margin) loss** \( z \mapsto \frac{1}{2}(1 - z)^2 \) is **1-strongly-convex**.
(e.g., second derivative is 1.)
Convexity of key losses.

**Logistic loss** $z \mapsto \ln(1 + \exp(-z))$ is **strictly convex**.
(e.g., verify that second derivative is positive.)

**Squared (margin) loss** $z \mapsto \frac{1}{2}(1 - z)^2$ is **1-strongly-convex**.
(e.g., second derivative is 1.)

Combined with our earlier linear prediction calculation, logistic regression and least squares are convex!
8. Convex optimization problems
Optimization problems

A typical optimization problem (in standard form) is written as

\[
\min_{x \in \mathbb{R}^d} \ f_0(x) \\
\text{s.t.} \ f_i(x) \leq 0 \quad \text{for all} \ i = 1, \ldots, n.
\]
A typical optimization problem (in standard form) is written as

$$\min_{x \in \mathbb{R}^d} f_0(x)$$

s.t.  $$f_i(x) \leq 0 \text{ for all } i = 1, \ldots, n.$$ 

▶ $f_0 : \mathbb{R}^d \to \mathbb{R}$ is the objective function;
A typical optimization problem (in standard form) is written as

$$\min_{x \in \mathbb{R}^d} f_0(x)$$

s.t. $$f_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, n.$$ 

- $f_0 : \mathbb{R}^d \to \mathbb{R}$ is the objective function;
- $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ are the constraint functions;
Optimization problems

A typical optimization problem (in standard form) is written as

$$\min_{x \in \mathbb{R}^d} f_0(x)$$

s.t. $$f_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, n.$$ 

- $f_0 : \mathbb{R}^d \to \mathbb{R}$ is the \textit{objective function};
- $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ are the \textit{constraint functions};
- inequalities $f_i(x) \leq 0$ are \textit{constraints};
A typical optimization problem (in standard form) is written as

$$\begin{align*}
\min_{x \in \mathbb{R}^d} & \quad f_0(x) \\
\text{s.t.} & \quad f_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, n.
\end{align*}$$

- $f_0 : \mathbb{R}^d \to \mathbb{R}$ is the \textit{objective function};
- $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ are the \textit{constraint functions};
- inequalities $f_i(x) \leq 0$ are \textit{constraints};
- $A := \left\{ x \in \mathbb{R}^d : f_i(x) \leq 0 \text{ for all } i = 1, 2, \ldots, n \right\}$ is the \textit{feasible region}. 
Optimization problems

A typical optimization problem (in standard form) is written as

\[
\min_{x \in \mathbb{R}^d} f_0(x) \\
\text{s.t. } f_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

- \( f_0 : \mathbb{R}^d \to \mathbb{R} \) is the objective function;
- \( f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R} \) are the constraint functions;
- inequalities \( f_i(x) \leq 0 \) are constraints;
- \( A := \{ x \in \mathbb{R}^d : f_i(x) \leq 0 \text{ for all } i = 1, 2, \ldots, n \} \) is the feasible region.
- **Goal**: Find \( x \in A \) so that \( f_0(x) \) is as small as possible.
Optimization problems

A typical optimization problem (in standard form) is written as

\[
\min_{x \in \mathbb{R}^d} \quad f_0(x)
\]

\[
\text{s.t.} \quad f_i(x) \leq 0 \quad \text{for all} \quad i = 1, \ldots, n.
\]

- \(f_0 : \mathbb{R}^d \to \mathbb{R}\) is the \textit{objective function};
- \(f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}\) are the \textit{constraint functions};
- inequalities \(f_i(x) \leq 0\) are \textit{constraints};
- \(A := \{x \in \mathbb{R}^d : f_i(x) \leq 0 \text{ for all } i = 1, 2, \ldots, n\}\) is the \textit{feasible region}.
- \textit{Goal}: Find \(x \in A\) so that \(f_0(x)\) is as small as possible.
- \textit{(Optimal) value} of the optimization problem is the smallest value \(f_0(x)\) achieved by a feasible point \(x \in A\).
A typical optimization problem (in standard form) is written as

\[
\min_{x \in \mathbb{R}^d} \quad f_0(x) \\
\text{s.t.} \quad f_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

- \( f_0 : \mathbb{R}^d \to \mathbb{R} \) is the objective function;
- \( f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R} \) are the constraint functions;
- inequalities \( f_i(x) \leq 0 \) are constraints;
- \( A := \{ x \in \mathbb{R}^d : f_i(x) \leq 0 \text{ for all } i = 1, 2, \ldots, n \} \) is the feasible region.
- **Goal**: Find \( x \in A \) so that \( f_0(x) \) is as small as possible.
- **(Optimal) value** of the optimization problem is the smallest value \( f_0(x) \) achieved by a feasible point \( x \in A \).
- Point \( x \in A \) achieving the optimal value is a (global) minimizer.
Convex optimization problems

Standard form of a *convex optimization problem*:

\[
\min_{x \in \mathbb{R}^d} \quad f_0(x) \\
\text{s.t.} \quad f_i(x) \leq 0 \quad \text{for all } i = 1, \ldots, n
\]

where \( f_0, f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R} \) are *convex functions*. 
Convex optimization problems

Standard form of a *convex optimization problem*:

\[
\begin{align*}
\min_{\mathbf{x} \in \mathbb{R}^d} & \quad f_0(\mathbf{x}) \\
\text{s.t.} & \quad f_i(\mathbf{x}) \leq 0 \quad \text{for all } i = 1, \ldots, n
\end{align*}
\]

where \( f_0, f_1, \ldots, f_n : \mathbb{R}^d \rightarrow \mathbb{R} \) are *convex functions*.

**Fact**: the feasible set

\[
A := \left\{ \mathbf{x} \in \mathbb{R}^d : f_i(\mathbf{x}) \leq 0 \text{ for all } i = 1, 2, \ldots, n \right\}
\]

is a convex set.

(SVMs next week will give us an example.)
Consider an optimization problem (not necessarily convex):

\[
\min_{x \in \mathbb{R}^d} f_0(x)
\]

s.t. \( x \in A \).

We say \( \tilde{x} \in A \) is a local minimizer if there is an "open ball" \( U := \{ x \in \mathbb{R}^d : \|x - \tilde{x}\|_2 < r \} \) of positive radius \( r > 0 \) such that \( \tilde{x} \) is a global minimizer for

\[
\min_{x \in \mathbb{R}^d} f_0(x)
\]

s.t. \( x \in A \cap U \).

Nothing looks better than \( \tilde{x} \) in the immediate vicinity of \( \tilde{x} \).

This is one local-to-global consequence of convexity; more generally, tangents lower bound the function everywhere.
Consider an optimization problem (not necessarily convex):

\[
\min_{x \in \mathbb{R}^d} f_0(x)
\]

s.t. \( x \in A \).

We say \( \tilde{x} \in A \) is a \textit{local minimizer} if there is an “open ball”

\[
U := \left\{ x \in \mathbb{R}^d : \| x - \tilde{x} \|_2 < r \right\}
\]

of positive radius \( r > 0 \) such that \( \tilde{x} \) is a global minimizer for

\[
\min_{x \in \mathbb{R}^d} f_0(x)
\]

s.t. \( x \in A \cap U \).
Consider an optimization problem (not necessarily convex):

$$\min_{x \in \mathbb{R}^d} f_0(x)$$

s.t. \( x \in A. \)

We say \( \tilde{x} \in A \) is a **local minimizer** if there is an “open ball”

$$U := \left\{ x \in \mathbb{R}^d : \|x - \tilde{x}\|_2 < r \right\}$$

of positive radius \( r > 0 \) such that \( \tilde{x} \) is a global minimizer for

$$\min_{x \in \mathbb{R}^d} f_0(x)$$

s.t. \( x \in A \cap U. \)

Nothing looks better than \( \tilde{x} \) **in the immediate vicinity** of \( \tilde{x} \).
Local minimizers

Consider an optimization problem (not necessarily convex):

\[
\min_{x \in \mathbb{R}^d} f_0(x) \\
\text{s.t.} \quad x \in A.
\]

We say \( \tilde{x} \in A \) is a **local minimizer** if there is an “open ball”

\[
U := \left\{ x \in \mathbb{R}^d : \| x - \tilde{x} \|_2 < r \right\}
\]

of positive radius \( r > 0 \) such that \( \tilde{x} \) is a global minimizer for

\[
\min_{x \in \mathbb{R}^d} f_0(x) \\
\text{s.t.} \quad x \in A \cap U.
\]

**Nothing looks better than \( \tilde{x} \) in the immediate vicinity of \( \tilde{x} \).**

This is one local-to-global consequence of convexity; more generally, tangents lower bound the function everywhere.
If the optimization problem is convex, and \( \tilde{x} \in A \) is a local minimizer, then it is also a global minimizer.
9. Convergence rates for gradient descent
Gradient descent

1. Let $w_0 \in \mathbb{R}^d$ be given.

2. For $i \in (0, 1, \ldots, t-1)$:
   2.1 $w_{i+1} := w_i - \eta_i \nabla f(w_i)$.

Intuition: convexity implies “no bumps”.
Smoothness

To analyze gradient descent, we'll use a notion of gradient stability.
Smoothness

To analyze gradient descent, we'll use a notion of gradient stability. \( \lambda \)-strong-convexity was a Taylor lower bound: \( \forall w, w', \)

\[
f(w') \geq f(w) + \nabla f(w)^\top (w' - w) + \frac{\lambda}{2} \| w' - w \|^2.
\]

Say \( f : \mathbb{R}^d \to \mathbb{R} \) is \( \beta \)-smooth when reverse holds: \( \forall w, w', \)

\[
f(w') \leq f(w) + \nabla f(w)^\top (w' - w) + \frac{\beta}{2} \| w' - w \|^2.
\]

(Also called Lipschitz gradients.)
**Theorem** (smoothness leads to approximate critical points). Let $w_0$ be given, and $w_{i+1} := w_i - \eta \nabla f(w_i)$. If $f$ is $\beta$-smooth and $\eta = 1/\beta$,

$$
\min_{i \leq t} \|\nabla f(w_{i-1})\|^2 \leq \frac{1}{t} \sum_{i=1}^{t} \|\nabla f(w_{i-1})\|^2 \leq \frac{2\beta}{t} \left( f(w_0) - \min_w f(w) \right).
$$
GD for smooth, non-convex functions.

**Theorem** (smoothness leads to approximate critical points).

Let $w_0$ be given, and $w_{i+1} := w_i - \eta \nabla f(w_i)$.

If $f$ is $\beta$-smooth and $\eta = \frac{1}{\beta}$,

$$\min_{i \leq t} \|\nabla f(w_{i-1})\|^2 \leq \frac{1}{t} \sum_{i=1}^{t} \|\nabla f(w_{i-1})\|^2 \leq \frac{2\beta}{t} \left( f(w_0) - \min_w f(w) \right).$$

**Proof.** Combining the definitions with choice of iterates gives (for each $i \leq t$)

$$f(w_i) \leq f(w_{i-1}) - \nabla f(w_{i-1})^\top (w_i - w_{i-1}) + \frac{\beta}{2} \|w_i - w_{i-1}\|^2$$

$$= f(w_{i-1}) - \frac{1}{2\beta} \|\nabla f(w_{i-1})\|^2.$$
Theorem (smoothness leads to approximate critical points).
Let $w_0$ be given, and $w_{i+1} := w_i - \eta \nabla f(w_i)$.
If $f$ is $\beta$-smooth and $\eta = \frac{1}{\beta}$,

$$\min_{i \leq t} \|\nabla f(w_{i-1})\|^2 \leq \frac{1}{t} \sum_{i=1}^{t} \|\nabla f(w_{i-1})\|^2 \leq \frac{2\beta}{t} \left(f(w_0) - \min_{w} f(w)\right).$$

Proof. Combining the definitions with choice of iterates gives (for each $i \leq t$)

$$f(w_i) \leq f(w_{i-1}) - \nabla f(w_{i-1})^\top(w_i - w_{i-1}) + \frac{\beta}{2} \|w_i - w_{i-1}\|^2$$

$$= f(w_{i-1}) - \frac{1}{2\beta} \|\nabla f(w_{i-1})\|^2.$$

Averaging these inequalities (over $i \leq t$) gives

$$\frac{1}{t} \sum_{i=1}^{t} \|\nabla f(w_{i-1})\|^2 \leq \frac{2\beta}{t} (f(w_0) - f(w_t)).$$
Theorem.
Let \( w_0 \) be given, and \( w_{i+1} := w_i - \eta \nabla f(w_i) \).
If convex \( f \) is \( \beta \)-smooth and \( \eta = 1/\beta \), \( \forall u \in \mathbb{R}^d \)

\[
f(w_t) - f(u) \leq \frac{1}{t} \sum_{i=1}^{t} (f(w_i) - f(u)) \leq \frac{\beta}{2t} \left( \|w_0 - u\|^2 - \|w_t - u\|^2 \right).
\]
GD for smooth, convex functions.

**Theorem.**
Let $w_0$ be given, and $w_{i+1} := w_i - \eta \nabla f(w_i)$.
If convex $f$ is $\beta$-smooth and $\eta = 1/\beta$, $\forall u \in \mathbb{R}^d$

$$f(w_t) - f(u) \leq \frac{1}{t} \sum_{i=1}^{t} (f(w_i) - f(u)) \leq \frac{\beta}{2t} \left( \|w_0 - u\|^2 - \|w_t - u\|^2 \right).$$

**Proof.** For each $i \leq t$, using the previous proof,

$$\|w_i - u\|^2 = \|w_{i-1} - u\|^2 - 2\eta \nabla f(w_{i-1})^\top (w_{i-1} - u) + \eta^2 \|\nabla f(w_{i-1})\|^2$$

$$\leq \|w_{i-1} - u\|^2 + 2\eta (f(u) - f(w_{i-1})) + 2\eta^2 \beta (f(w_{i-1}) - f(w_i))$$

$$= \|w_{i-1} - u\|^2 + \frac{2}{\beta} (f(u) - f(w_i)).$$

Rearranging and then averaging these inequalities over $i \leq t$ gives the bound.
10. Convexity and differentiability
Many useful convex functions are *not differentiable*.

\[ x \mapsto |x|. \]

**Question:** how can we do gradient descent?
Subgradients

Derivatives give tangents and descent directions:
\[ f(w') \geq f(x) + \nabla f(x)^T (w' - x). \]

Subdifferential set:
\[ \partial f(x) = \left\{ s \in \mathbb{R}^d : \forall w'. \ f(w') \geq f(x) + s^T (w' - x) \right\}. \]
Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex.

**First order conditions:** For any $y \in \mathbb{R}^d$,

$$0 \in \partial f(y) \iff f(y) = \inf_{x} f(x).$$
Subgradients: first order condition.

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex.

**First order conditions:** For any $y \in \mathbb{R}^d$,

$$0 \in \partial f(y) \iff f(y) = \inf_x f(x).$$

**Magic of convexity:** local information gives global structure.
If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then $\mathbb{E}f(X) \geq f(\mathbb{E}X)$.

Proof. Set $y := \mathbb{E}X$, and pick any $s \in \partial f(\mathbb{E}X)$. Then

$$
\mathbb{E}f(X) \geq \mathbb{E}\left(f(y) + s^T(X - y)\right) = f(y) + s^T\mathbb{E}(X - y) = f(y).
$$

Note. This inequality comes up often!
11. Summary
Summary

- Convex sets and functions.
- Ways to verify convexity.
- Strict convexity, strong convexity.
- Intuition for gradient descent convergence: local-to-global structure, no bumps.
- Jensen’s inequality.