Convex optimization

CS 446 / ECE 449

2022-02-01 04:33:47 -0600 (`6fe12f5`)

Why are we learning about convexity?

We've already seen it twice:

- If $\widehat{\mathcal{R}}$ is convex, then any $\boldsymbol{w}$ with $\nabla \widehat{\mathcal{R}}(\boldsymbol{w})$ is globally optimal: $\widehat{\mathcal{R}}(\boldsymbol{w}) = \inf_{\boldsymbol{v}} \widehat{\mathcal{R}}(\boldsymbol{v})$. Matches the normal equations!
- Gradient descent on convex $\widehat{\mathcal{R}}$ is guaranteed to work.

Convexity is pervasive throughout mathematics.

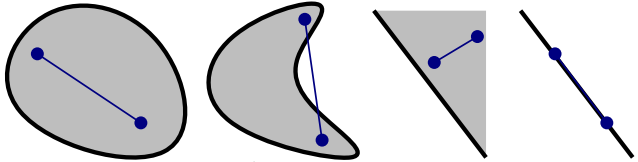
Warning: $\widehat{\mathcal{R}}$ is **not** convex for deep networks.

Plan for today

- Convex sets and functions.
- Minimizing convex functions.

Convex sets

A set S is convex if, for every pair of points $\{\boldsymbol{x}, \boldsymbol{x}'\}$ in S, the line segment between $\boldsymbol{x}$ and $\boldsymbol{x}'$ is also contained in S.

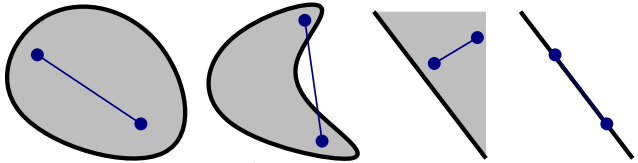


In symbols:

$$\{\boldsymbol{x}, \boldsymbol{x}'\} \in S \quad \Longrightarrow \quad [\boldsymbol{x}, \boldsymbol{x}'] = \{\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{x}' : \alpha \in [0,1]\} \subseteq S.$$

Convex sets

A set S is convex if, for every pair of points $\{\boldsymbol{x}, \boldsymbol{x}'\}$ in S, the line segment between $\boldsymbol{x}$ and $\boldsymbol{x}'$ is also contained in S.



In symbols:

$$\{\boldsymbol{x}, \boldsymbol{x}'\} \in S \qquad \Longrightarrow \qquad [\boldsymbol{x}, \boldsymbol{x}'] = \{\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}' : \alpha \in [0,1]\} \subseteq S.$$

Examples:

- All of $\mathbb{R}^d$.
- Empty set.
- Half-spaces: $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}^\mathsf{T}\boldsymbol{x} \leq b\}$.
- Intersections of convex sets.
- Polyhedra: $\left\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\right\} = \bigcap_{i=1}^m \left\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}_i^\mathsf{T}\boldsymbol{x} \leq b_i\right\}$.
- Convex hulls: $\mathrm{conv}(S) := \{\sum_{i=1}^k \alpha_i \boldsymbol{x}_i : k \in \mathbb{N},\, \boldsymbol{x}_i \in S,\, \alpha_i \geq 0,\, \sum_{i=1}^k \alpha_i = 1\}$.

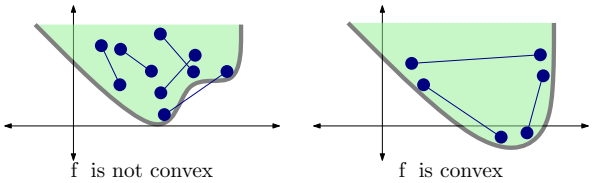
Let's verify that any halfspace $H := \left\{ \boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}^\top \boldsymbol{x} \leq b \right\}$ is convex.

Convex functions from convex sets

The epigraph of a function f is the region above the curve:

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

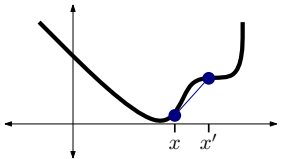
A function is convex if its epigraph is a convex set.



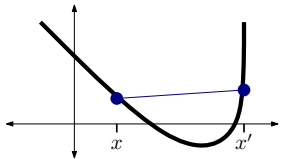
Convex functions (standard definitions)

(Secants lie above.) $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex if for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $\alpha \in [0,1]$,

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



f is not convex

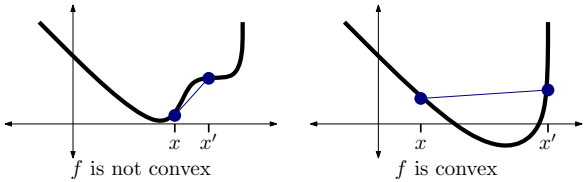


f is convex

Convex functions (standard definitions)

(Secants lie above.) $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex if for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $\alpha \in [0,1]$,

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



f is not convex $\qquad$ f is convex

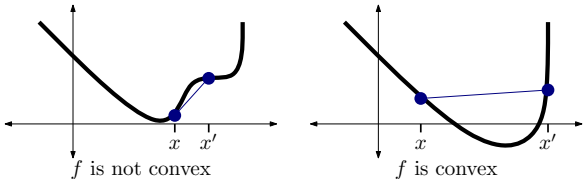
Other characterizations.

- **(Tangents lie below.)** Given $\boldsymbol{x}, \boldsymbol{z}$, then $f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{z} - \boldsymbol{x})$.
- **(Increasing slopes.)** Given $\boldsymbol{x}, \boldsymbol{z}$, then $\left(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{z})\right)^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{z}) \geq 0$.
- **(Curves upwards.)** Given $\boldsymbol{x}$, then $\nabla^2 f(\boldsymbol{x}) \succeq 0$.

Convex functions (standard definitions)

(Secants lie above.) $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex if for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and $\alpha \in [0,1]$,

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



Other characterizations.

- **(Tangents lie below.)** Given $\boldsymbol{x}, \boldsymbol{z}$, then $f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{z}-\boldsymbol{x})$.
- **(Increasing slopes.)** Given $\boldsymbol{x}, \boldsymbol{z}$, then $\left(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{z})\right)^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{z}) \geq 0$.
- **(Curves upwards.)** Given $\boldsymbol{x}$, then $\nabla^2 f(\boldsymbol{x}) \succeq 0$.

Examples.

- $f(\boldsymbol{x}) = \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x}$ for any $\boldsymbol{b} \in \mathbb{R}^d$.
- $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}$ for symmetric positive semidefinite $\boldsymbol{A}$.
- $f(\boldsymbol{x}) = \ln\left(\sum_{i=1}^d \exp(x_i)\right)$, which approximates $\max_i x_i$.

Is $f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{M}\boldsymbol{x}$ convex? Does the answer depend on dimension?

Operations preserving convexity

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

$$\boldsymbol{x} \mapsto \alpha_1 f_1(\boldsymbol{x}) + \cdots + \alpha_k f_k(\boldsymbol{x}) \qquad \text{is convex.}$$

Affine composition: if f is convex, then for any $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{b} \in \mathbb{R}^m$,

$$\boldsymbol{x} \mapsto f\left(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\right) \qquad \text{is convex.}$$

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

$$\boldsymbol{x} \mapsto \max_i f_i(\boldsymbol{x}) \qquad \text{is convex.}$$

Examples: losses and empirical risks.

Our standard losses are convex.

- **(Logistic loss.)** $\ell_{\log}(z) := \ln(1 + \exp(-z))$ is convex.
- **(Squared loss.)** $\ell_1(z) := \frac{1}{2}(1 - z)^2$ is convex in z, $\ell_2(y, \hat{y}) := \frac{1}{2}(y - \hat{y})^2$ is convex in $\hat{y}$.

Examples: losses and empirical risks.

Our standard losses are convex.

- **(Logistic loss.)** $\ell_{\log}(z) := \ln(1 + \exp(-z))$ is convex.
- **(Squared loss.)** $\ell_1(z) := \frac{1}{2}(1 - z)^2$ is convex in z,
  $\ell_2(y, \hat{y}) := \frac{1}{2}(y - \hat{y})^2$ is convex in $\hat{y}$.

Empirical risks.

- If ℓ is convex over $\mathbb{R}$, then $\widehat{\mathcal{R}}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \ell(y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w})$ is convex over $\mathbb{R}^d$.
- If ℓ is convex, and we use a nonlinear predictor $f_{\boldsymbol{w}} : \mathbb{R}^d \to \mathbb{R}$,
  then $\widehat{\mathcal{R}}(f_{\boldsymbol{w}}) := \frac{1}{n}\sum_{i=1}^n \ell(y_i f_{\boldsymbol{w}}(\boldsymbol{x}_i))$ may be non-convex.

Convexity and optimization

Gradient descent behaves nicely for convex functions.

Many principles of its behavior seem to carry over to deep network optimization.

First-order conditions

Fact.
If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, then $\boldsymbol{x}$ is globally optimal iff $\nabla f(\boldsymbol{x}) = 0$.

Implication: $\boldsymbol{w}$ satisfies the normal equations iff $\boldsymbol{w}$ is globally optimal for $\widehat{\mathcal{R}}_{\mathsf{sq}}$.

Differentiability

Many convex functions are not differentiable
(e.g., $\boldsymbol{x} \mapsto \max_i x_i$, or $z \mapsto |z|$, or $\mathrm{ReLU}(z) := \max\{0, z\}$.)

(Tangents lie below.) Given $f : \mathbb{R}^d \to \mathbb{R}$ and $\boldsymbol{x}$, subdifferential set $\partial f(\boldsymbol{x})$ is

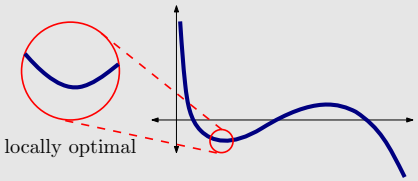
$$\partial f(\boldsymbol{x}) := \left\{ \boldsymbol{s} \in \mathbb{R}^d \ : \ \forall \boldsymbol{z} \in \mathbb{R}^d, f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{s}^{\mathsf{T}}(\boldsymbol{z} - \boldsymbol{x}) \right\}.$$

Properties.

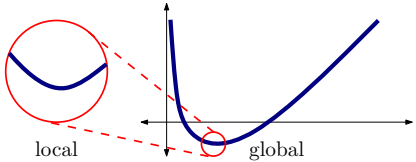
- If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then ∂f is nonempty everywhere.
- $\boldsymbol{x}$ globally optimizes f iff $0 \in \partial f(\boldsymbol{x})$.
- Gives an easy proof of Jensen's inequality: $f(\mathbb{E}X) \leq \mathbb{E}f(X)$ for convex f.

Behavior of gradient descent

Recall that gradient descent rolls down hills: $\boldsymbol{x}' := \boldsymbol{x} - \eta \nabla f(\boldsymbol{x})$.



Without convexity,
we only get stationary points.



With convexity,
we get global optima.

Local-to-global principle: both cases pass near stationary points, but for convex functions this implies optimality.

Convergence rates for gradient descent

Additional assumption: suppose $f : \mathbb{R}^d \to \mathbb{R}$ is **"β-smooth"**: for any $\boldsymbol{w}, \boldsymbol{w}'$,

$$f(\boldsymbol{w}') \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\mathsf{T}(\boldsymbol{w}' - \boldsymbol{w}) + \frac{\beta}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2.$$

(In words: f never curves upwards faster than a quadratic.)

Convergence rates for gradient descent

Additional assumption: suppose $f : \mathbb{R}^d \to \mathbb{R}$ is **"β-smooth"**: for any $\boldsymbol{w}, \boldsymbol{w}'$,

$$f(\boldsymbol{w}') \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\mathsf{T}(\boldsymbol{w}' - \boldsymbol{w}) + \frac{\beta}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2.$$

(In words: f never curves upwards faster than a quadratic.)

Gradient descent convergence rates for β-smooth f with $\eta := \frac{1}{\beta}$.

1. Without convexity,

$$\min_{i \leq t} \left\|\nabla f(\boldsymbol{w}_{i-1})\right\|^2 \leq \frac{2\beta}{t}\left(f(\boldsymbol{w}_0) - \inf_{\boldsymbol{w}} f(\boldsymbol{w})\right).$$

2. With convexity, for any $\boldsymbol{u} \in \mathbb{R}^d$,

$$f(\boldsymbol{w}_t) - f(\boldsymbol{u}) \leq \frac{\beta}{2t}\left(\|\boldsymbol{w}_0 - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{u}\|^2\right).$$

Proofs are in appendix.

Stochastic gradients

Gradient requires time $\mathcal{O}(n)$:

$$\nabla \widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i f(\boldsymbol{x}_i; \boldsymbol{w})) y_i \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_i; \boldsymbol{w}).$$

Instead use a stochastic minibatch S of size $|S| = B$:

$$\frac{1}{B} \sum_{i \in S} \ell'(y_i f(\boldsymbol{x}_i; \boldsymbol{w})) y_i \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_i; \boldsymbol{w}).$$

- In the convex case, can still prove convergence rates.
- Increasing B improves gradient accuracy;
  in practice, hardware issues like memory size dictate B.
  Common choices in deep learning are $[256, 2048]$.

Other practical considerations

There are many other gradient-based ("first-order") methods.
`torch.optim` contains many variants (e.g., Adam).

Other practical considerations

There are many other gradient-based ("first-order") methods.
`torch.optim` contains many variants (e.g., Adam).

Deep networks are not differentiable.
They are not convex, so we can't even use subgradients.
Pytorch does not correctly compute gradients, even for convex functions.
No one cares, since `torch.optim` seems to work.

Convex programs

A **convex program** is minimization of a convex function f over a convex set S:

$$\min_{\boldsymbol{x} \in S} f(\boldsymbol{x}).$$

- Gradient methods can be adjusted to work well here.
- This used to be a big research area, but deep learning doesn't make any serious use of this formulation or its tools.

Summary for today

- Convex sets and functions.
- Minimizing convex functions.

(Appendix.)

(Didn't mention: f is concave means $-f$ is convex.)

Convex functions can lack curvature: $x \mapsto x$ and $x \mapsto |x|$ are convex.
There are two standard notions of curved convex functions:

- strict convexity: no flat regions;
- strong convexity: more curved than a quadratic, everywhere.

Strict convexity

Function values: $\forall \boldsymbol{x}, \boldsymbol{y}$, $\forall \alpha \in [0,1]$:

$$f\left(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\right) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}).$$

Derivatives: $\forall \boldsymbol{x}, \boldsymbol{y}$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

Hessians: $\forall \boldsymbol{x}$,

$$\nabla^2 f(\boldsymbol{x}) \succeq 0.$$

Strict convexity

Function values: $\forall \boldsymbol{x} \neq \boldsymbol{y}$, $\forall \alpha \in (0,1)$:

$$f\left(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\right) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}).$$

Derivatives: $\forall \boldsymbol{x} \neq \boldsymbol{y}$,

$$f(\boldsymbol{y}) > f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

Hessians: $\forall \boldsymbol{x}$,

$$\nabla^2 f(\boldsymbol{x}) \succ 0.$$

λ-Strong-Convexity.

Function values: $\forall \boldsymbol{x}, \boldsymbol{y}$, $\forall \alpha \in [0,1]$

$$f\left(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\right) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}).$$

Derivatives: $\forall \boldsymbol{x}, \boldsymbol{y}$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\top}(\boldsymbol{y} - \boldsymbol{x}).$$

Hessians: $\forall \boldsymbol{x}$,

$$\nabla^2 f(\boldsymbol{x}) \succeq 0.$$

λ-Strong-Convexity.

Function values: $\forall \boldsymbol{x}, \boldsymbol{y}$, $\forall \alpha \in [0,1]$

$$f\left(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\right) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - \frac{\lambda \alpha (1-\alpha)}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Derivatives: $\forall \boldsymbol{x}, \boldsymbol{y}$

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

Hessians: $\forall \boldsymbol{x}$,

$$\nabla^2 f(\boldsymbol{x}) \succeq \lambda \boldsymbol{I}.$$

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 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 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!

GD for smooth, non-convex functions.

Theorem (smoothness implies stationary points).

GD for smooth, non-convex functions.

Theorem (smoothness implies stationary points).
Let $\boldsymbol{w}_0$ be given, and $\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta \nabla f(\boldsymbol{w}_i)$.
If f is β-smooth and $\eta = 1/\beta$,

$$\min_{i \leq t} \left\| \nabla f(\boldsymbol{w}_{i-1}) \right\|^2 \leq \frac{1}{t} \sum_{i=1}^{t} \left\| \nabla f(\boldsymbol{w}_{i-1}) \right\|^2 \leq \frac{2\beta}{t} \left(f(\boldsymbol{w}_0) - \min_{\boldsymbol{w}} f(\boldsymbol{w}) \right).$$

GD for smooth, non-convex functions.

Theorem (smoothness implies stationary points).
Let $\boldsymbol{w}_0$ be given, and $\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta \nabla f(\boldsymbol{w}_i)$.
If f is β-smooth and $\eta = 1/\beta$,

$$\min_{i \le t} \left\| \nabla f(\boldsymbol{w}_{i-1}) \right\|^2 \le \frac{1}{t} \sum_{i=1}^{t} \left\| \nabla f(\boldsymbol{w}_{i-1}) \right\|^2 \le \frac{2\beta}{t} \left(f(\boldsymbol{w}_0) - \min_{\boldsymbol{w}} f(\boldsymbol{w}) \right).$$

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

$$\begin{aligned} f(\boldsymbol{w}_i) &\le f(\boldsymbol{w}_{i-1}) + \nabla f(\boldsymbol{w}_{i-1})^\top (\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) + \frac{\beta}{2} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|^2 \\ &= f(\boldsymbol{w}_{i-1}) - \frac{1}{2\beta} \left\| \nabla f(\boldsymbol{w}_{i-1}) \right\|^2. \end{aligned}$$

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

$$\frac{1}{t} \sum_{i=1}^{t} \left\| \nabla f(\boldsymbol{w}_{i-1}) \right\|^2 \le \frac{2\beta}{t} \left(f(\boldsymbol{w}_0) - f(\boldsymbol{w}_t) \right).$$

□

GD for smooth, convex functions.

Theorem (smoothness and convexity imply global optimality.

GD for smooth, convex functions.

Theorem (smoothness and convexity imply global optimality.
Let $\boldsymbol{w}_0$ be given, and $\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta \nabla f(\boldsymbol{w}_i)$.
If f is convex and β-smooth, and $\eta = 1/\beta$, then $\forall \boldsymbol{u} \in \mathbb{R}^d$,

$$f(\boldsymbol{w}_t) - f(\boldsymbol{u}) \leq \frac{1}{t} \sum_{i=1}^{t} \big(f(\boldsymbol{w}_i) - f(\boldsymbol{u})\big) \leq \frac{\beta}{2t} \left(\|\boldsymbol{w}_0 - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{u}\|^2\right).$$

GD for smooth, convex functions.

Theorem (smoothness and convexity imply global optimality.
Let $\boldsymbol{w}_0$ be given, and $\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta \nabla f(\boldsymbol{w}_i)$.
If f is convex and β-smooth, and $\eta = 1/\beta$, then $\forall \boldsymbol{u} \in \mathbb{R}^d$,

$$f(\boldsymbol{w}_t) - f(\boldsymbol{u}) \leq \frac{1}{t} \sum_{i=1}^{t} \big(f(\boldsymbol{w}_i) - f(\boldsymbol{u})\big) \leq \frac{\beta}{2t} \left(\|\boldsymbol{w}_0 - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_t - \boldsymbol{u}\|^2\right).$$

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

$$\begin{aligned}
\|\boldsymbol{w}_i - \boldsymbol{u}\|^2 &= \|\boldsymbol{w}_{i-1} - \boldsymbol{u}\|^2 - 2\eta \nabla f(\boldsymbol{w}_{i-1})^\top (\boldsymbol{w}_{i-1} - \boldsymbol{u}) + \eta^2 \big\|\nabla f(\boldsymbol{w}_{i-1})\big\|^2 \\
&\leq \|\boldsymbol{w}_{i-1} - \boldsymbol{u}\|^2 + 2\eta \big(f(\boldsymbol{u}) - f(\boldsymbol{w}_{i-1})\big) + 2\eta^2 \beta \big(f(\boldsymbol{w}_{i-1}) - f(\boldsymbol{w}_i)\big) \\
&= \|\boldsymbol{w}_{i-1} - \boldsymbol{u}\|^2 + \frac{2}{\beta} \big(f(\boldsymbol{u}) - f(\boldsymbol{w}_i)\big).
\end{aligned}$$

Rearranging and then averaging these inequalities over $i \leq t$ gives the bound. □

Example: OLS

As a consequence, for least squares, starting from $\boldsymbol{w}_0 = 0$,

$$\widehat{\mathcal{R}}(\boldsymbol{w}_t) - \widehat{\mathcal{R}}(\hat{\boldsymbol{w}}_{\text{ols}}) \leq \frac{\beta \|\hat{\boldsymbol{w}}_{\text{ols}}\|^2}{2t},$$

where β and $\|\hat{\boldsymbol{w}}_{\text{ols}}\|$ depend on $\boldsymbol{X}$ and $\boldsymbol{y}$.

Jensen's inequality

Jensen's inequality: if $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then $\mathbb{E}f(\boldsymbol{X}) \geq f(\mathbb{E}\boldsymbol{X})$.
Proof. Set $\boldsymbol{y} := \mathbb{E}\boldsymbol{X}$, and pick any $\boldsymbol{s} \in \partial f(\mathbb{E}\boldsymbol{X})$. Then

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

Supplemental reading

- Shalev-Shwartz/Ben-David: chapter 12.
- Further advanced reading on convexity and convex optimization:
  - Convex optimization: course lecture slides by Lieven Vandenberghe part 1 part 2 part 3, "introductory lectures on convex optimization" by Nesterov.
  - Convexity: "fundamentals of convex analysis" by Hiriart-Urruty and Lemaréchal, "convex analysis and nonlinear optimization" by Borwein and Lewis.