Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

L4: Optimization Primal.

Note: all figures re-done on board in lecture.

Lecture outline.

1. Review.
2. Convexity.
3. Descent methods.

Reading.

- **Convexity and optimization:** Boyd and Vandenberghe “Convex Optimization”, Chapters 2-4.
- **Convexity** (*very optional! beyond this course.*)**:** Hiriart-Urruty and Lemaréchal, “Fundamentals of Convex Analysis”; Borwein and Lewis, “Convex Analysis and Nonlinear Optimization”.

Linear classification.

Suppose $y^{(i)} \in \{-1, +1\}$.
ERM (Empirical Risk Minimization) for linear classification:

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}\right) \qquad \text{or} \qquad \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y^{(i)}\boldsymbol{w}^\top \phi(\mathbf{x}^{(i)})\right).$$

Some choices for loss ℓ:

$$\begin{aligned}
&z \mapsto \mathbf{1}[z \le 0] && \text{zero-one/classification,}\\
&z \mapsto \frac{1}{2}(1-z)^2 && \text{least squares/linear regression,}\\
&z \mapsto \ln(1+\exp(-z)) && \text{logistic.}
\end{aligned}$$

Linear classification.

Some choices for loss ℓ:

$$
\begin{aligned}
z &\mapsto \mathbf{1}[z \le 0] && \text{zero-one/classification,} \\
z &\mapsto \frac{1}{2}(1 - z)^2 && \text{least squares/linear regression,} \\
z &\mapsto \ln(1 + \exp(-z)) && \text{logistic.}
\end{aligned}
$$

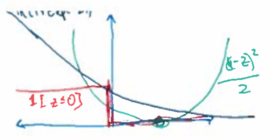
Linear classification.

Some choices for loss ℓ:

$$z \mapsto \mathbf{1}[z \leq 0] \qquad \text{zero-one/classification,}$$

$$z \mapsto \frac{1}{2}(1 - z)^2 \qquad \text{least squares/linear regression,}$$

$$z \mapsto \ln(1 + \exp(-z)) \qquad \text{logistic.}$$



Descent methods.

How to solve for **weights** $\boldsymbol{w} \in \mathbb{R}^d$ in

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}) \ ?$$

Can use **gradient descent.** But why should it work?

Descent methods.

How to solve for **weights** $\boldsymbol{w} \in \mathbb{R}^d$ in

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) \ ?$$

Can use **gradient descent.** But why should it work?

Today: linear and logistic regression have no “bumps”.

Descent methods and neural networks.

Neural nets are not linear models, but still have **weights**:

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) \qquad \text{linear predictor,}$$

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)} f_{\boldsymbol{w}}(\mathbf{x}^{(i)})) \qquad \text{neural net.}$$

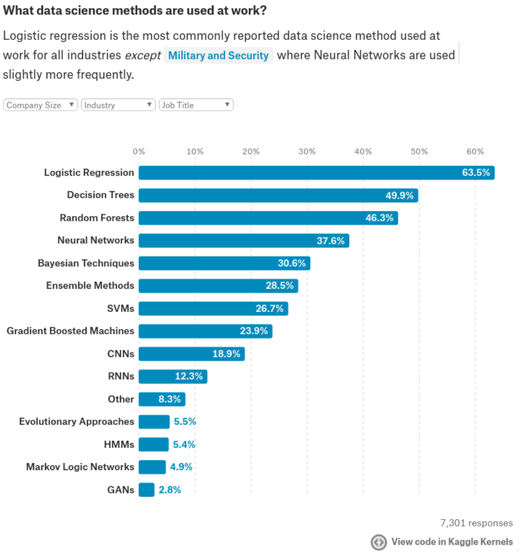
Encounter “bumps” with neural nets?

Kaggle survey.

Lastly: is logistic regression relevant?

Kaggle survey.

Lastly: is logistic regression relevant?



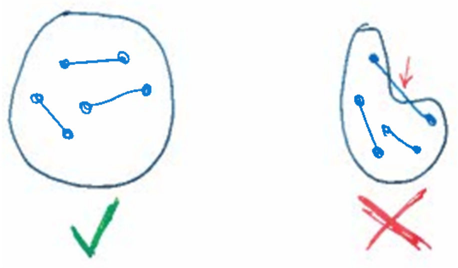
Convexity.

Convexity.

Convexity will formalize “no bumps”.
Convexity is pervasive in mathematics, not just optimization.

Convex sets.

A set is **convex** if it contains all line segments:



In symbols: $C \subseteq \mathbb{R}^d$ is convex when

$$\{\boldsymbol{x}, \boldsymbol{y}\} \subseteq C \qquad \Longrightarrow \qquad [\boldsymbol{x}, \boldsymbol{y}] \subseteq C$$

where $[\boldsymbol{x}, \boldsymbol{y}] = \big\{\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y} : \alpha \in [0, 1]\big\}$.

Convex set operations.

Convex hull is similar to putting a rubber band around data:

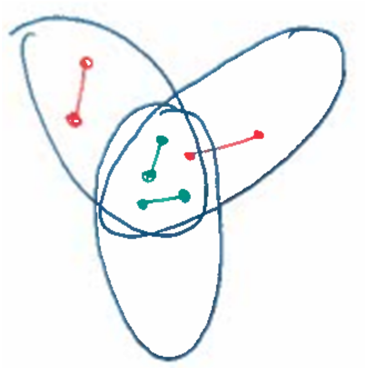
Rigorously: intersection of all convex supersets.

Alternatively (for finite set $S = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$):
all **convex combinations**:

$$\operatorname{conv}(S) := \left\{ \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i \; : \; \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1 \right\}.$$

Convex set intersections.

Convex hull is the intersection of all convex supersets. . .
. . . why is convexity preserved under intersection?

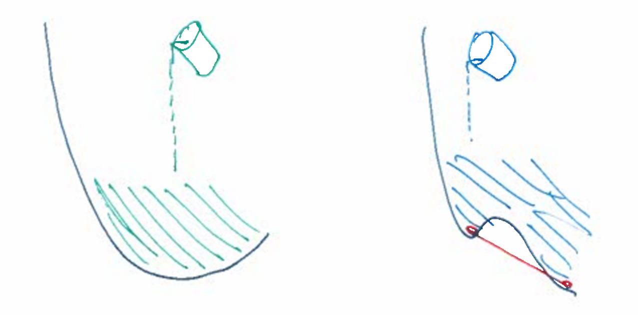


Example. Polyhedron $\left\{ \boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{Ax} \leq \mathbf{b} \right\}$.

Convex functions.

Bucket fill a function from above: this is the **epigraph**

$$\operatorname{epi}(f) := \left\{ (\boldsymbol{x}, r) \;:\; \boldsymbol{x} \in \mathbb{R}^d, r \in \mathbb{R}, f(\boldsymbol{x}) \le r \right\}.$$

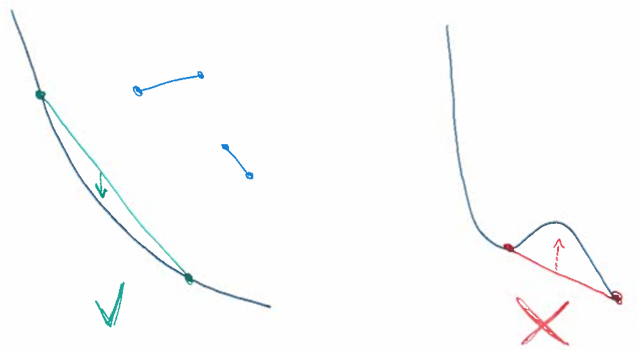


f is a **convex function** when $\operatorname{epi}(f)$ is a convex set.

Convex functions — algebraic form.

Equivalently: for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\alpha \in [0, 1]$,

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



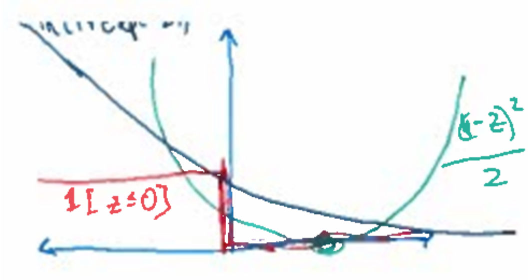
Convex functions — no bumps!

Gradient descent works when no “bumps”.

Convexity implies no bumps!

Convex losses.

The logistic and least squares losses *look convex*:



Question.

(a) How to prove this?

(b) What about convexity of *risk* $\boldsymbol{w} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell\left(y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)}\right)$?

Other convex functions.

- Exponential: e^x.
- Negative logarithm: $-\log(x)$ over $\mathbb{R}_{>0}$.
- Negative entropy: $x\log(x)$ over $\mathbb{R}_{>0}$.
- Norms: $\|\boldsymbol{x}\|_p$ for $p \geq 1$.
- Log-Sum-Exp: $\ln\big(\exp(\boldsymbol{x}_1) + \cdots + \exp(\boldsymbol{x}_d)\big)$.

Three checks for convexity.

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

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

Three checks for convexity.

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

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

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

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

(This implies *increasing slopes*: $\left(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right)^\top (\boldsymbol{y} - \boldsymbol{x}) \geq 0$.)

Three checks for convexity.

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

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

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

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

(This implies *increasing slopes*: $\left(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right)^\top (\boldsymbol{y} - \boldsymbol{x}) \geq 0$.)

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

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

(We'll use this for least squares and logistic losses.)

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.**

Squared loss $z \mapsto \frac{1}{2}(1 - z)^2$ is 1**-strongly-convex**.

What about the **risk** $\boldsymbol{w} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)})$?

Operations preserving convexity.

If $(f_1, \ldots, f_k)$ convex and $(\alpha_1, \ldots, \alpha_k)$ nonnegative,

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

If f is convex, then for any matrix $\boldsymbol{A}$ and vector $\mathbf{b}$,

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

If $\mathcal{F}$ is a *set* of convex functions,

$$\boldsymbol{w} \mapsto \sup_{f \in \mathcal{F}} f(\boldsymbol{w}) \qquad \text{is convex.}$$

Convexity of linear prediction with convex losses.

Suppose loss $\ell : \mathbb{R} \to \mathbb{R}$ is convex.
Collect $y^{(i)}\mathbf{x}^{(i)}$ into a single matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$:

$$\boldsymbol{A} := \begin{bmatrix} \longleftarrow & y^{(1)}\mathbf{x}^{(1)} & \longrightarrow \\ & \vdots & \\ \longleftarrow & y^{(n)}\mathbf{x}^{(n)} & \longrightarrow \end{bmatrix}.$$

Then $\boldsymbol{v} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{v}_i)$ is convex, as is

$$\boldsymbol{w} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^{n} \frac{1}{n}\ell\left((\boldsymbol{A}w)_i\right),$$

Alternatively:

$$\boldsymbol{w} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^{n} \frac{1}{n}\ell\left(\boldsymbol{w}^\top (y^{(i)}\mathbf{x}^{(i)})\right).$$

Convexity of linear prediction with convex losses.

Suppose loss $\ell : \mathbb{R} \to \mathbb{R}$ is convex.
Alternatively:

$$\boldsymbol{w} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^{n} \frac{1}{n}\ell\left(\boldsymbol{w}^\top (y^{(i)}\mathbf{x}^{(i)})\right).$$

Convexity of linear prediction with convex losses.

Suppose loss $\ell : \mathbb{R} \to \mathbb{R}$ is convex.
Alternatively:

$$\boldsymbol{w} \mapsto \frac{1}{n} \sum_{i=1}^{n} \ell(y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^{n} \frac{1}{n} \ell \left(\boldsymbol{w}^\top (y^{(i)} \mathbf{x}^{(i)}) \right).$$

Therefore linear and logistic regression are convex minimization!

Convexity of linear prediction with convex losses.

Suppose loss $\ell : \mathbb{R} \to \mathbb{R}$ is convex.
Alternatively:

$$\boldsymbol{w} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) = \sum_{i=1}^{n} \frac{1}{n}\ell\left(\boldsymbol{w}^\top (y^{(i)}\mathbf{x}^{(i)})\right).$$

Therefore linear and logistic regression are convex minimization!
Therefore gradient descent should work!

Convexity: subgradients.

Convexity and differentiability.

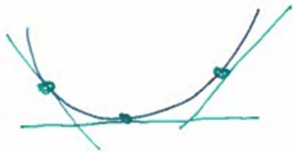
Many useful convex functions are *not differentiable*.

$$x \mapsto |x|.$$

Question: how can we do gradient descent?

Subgradients.

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

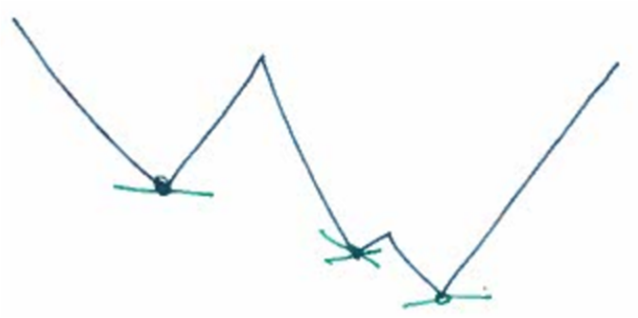


Subdifferential *set*:

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

Aside: neural network subdifferentials.

Standard neural network packages (tensorflow, pytorch, etc.) give weird “descent directions” for things without even subgradients



Subgradients: first order condition.

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex.
First order conditions: For any $\boldsymbol{y} \in \mathbb{R}^d$,

$$0 \in \partial f(\boldsymbol{y}) \qquad \iff \qquad f(\boldsymbol{y}) = \inf_{\boldsymbol{x}} f(\boldsymbol{x}).$$

Proof. By definition of subgradient!

Subgradients: first order condition.

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex.
First order conditions: For any $\boldsymbol{y} \in \mathbb{R}^d$,

$$0 \in \partial f(\boldsymbol{y}) \qquad \Longleftrightarrow \qquad f(\boldsymbol{y}) = \inf_{\boldsymbol{x}} f(\boldsymbol{x}).$$

Proof. By definition of subgradient!

Magic of convexity: local information gives global structure.

Subgradients: 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

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

Note. This inequality comes up often!

Further topics.

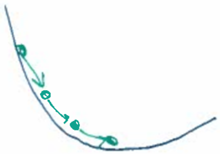
If you like this material,
e.g., you'd like to see another reason why $\frac{1}{2}\|\cdot\|^2$ has "$1/2$", see

- Hiriart-Urruty and Lemaréchal, "Fundamentals of Convex Analysis";
- Borwein and Lewis, "Convex Analysis and Nonlinear Optimization".

Descent methods.

Gradient descent.

1. Let $\boldsymbol{w}_0 \in \mathbb{R}^d$ be given.
2. For $i \in (0, 1, \ldots, t)$:
    1. $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.



Note. Can relax $\nabla f(w_{i-1})$ in various ways.

Smoothness.

λ**-strong-convexity** was a Taylor lower bound: $\forall \boldsymbol{x}, \boldsymbol{y}$,

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

Say $f : \mathbb{R}^d \to \mathbb{R}$ is β**-smooth** when reverse holds: $\forall \boldsymbol{x}, \boldsymbol{y}$,

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

Smooth, non-convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.
If f is β-smooth and $\alpha_i = {}^1/_\beta$,

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

Smooth, non-convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.
If f is β-smooth and $\alpha_i = {}^1\!/\!_\beta$,

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

Proof. Averaging the inequalities (for each $i \leq t$)

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

gives

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

Smooth, convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.
If convex f is β-smooth and $\alpha_i = 1/\beta$, for every $\mathbf{u} \in \mathbb{R}^d$

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

Smooth, convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.
If convex f is β-smooth and $\alpha_i = 1/\beta$, for every $\mathbf{u} \in \mathbb{R}^d$

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

Proof. Average the inequalities (for each $i \leq t$)

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

Smooth and strongly convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.

If f is β-smooth and λ-strongly-convex, $\alpha_i = 1/\beta$, and $\mathbf{u}$ is optimal,

$$f(\boldsymbol{w}_t) - f(\mathbf{u}) \leq \exp\left(-t\lambda/\beta\right).$$

Smooth and strongly convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.

If f is β-smooth and λ-strongly-convex, $\alpha_i = 1/\beta$, and $\mathbf{u}$ is optimal,

$$f(\boldsymbol{w}_t) - f(\mathbf{u}) \leq \exp\left(-t\lambda/\beta\right).$$

Proof. Since

$$\begin{aligned} f(\mathbf{u}) &= \inf_{\boldsymbol{v}} f(\boldsymbol{v}) = \inf_{\boldsymbol{v}} f(\boldsymbol{w}_i + \boldsymbol{v}) \\ &\geq \inf_{\boldsymbol{v}} \left(f(\boldsymbol{w}_i) + \nabla f(\boldsymbol{w}_i)^\top \boldsymbol{v} + \frac{\lambda}{2} \|\boldsymbol{v}\|^2 \right) \qquad = f(\boldsymbol{w}_i) - \frac{1}{2\lambda} \|\boldsymbol{w}_i\|^2, \end{aligned}$$

then

$$f(\boldsymbol{w}_i) \leq f(\boldsymbol{w}_{i-1}) - \frac{1}{2\beta} \left\|\nabla f(\boldsymbol{w}_{i-1})\right\|^2 \leq f(\boldsymbol{w}_{i-1}) - \frac{\lambda}{\beta} \left(f(\boldsymbol{w}_{i-1}) - f(\mathbf{u}) \right);$$

recurse.

Smooth and strongly convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.

If f is β-smooth and λ-strongly-convex, $\alpha_i = 1/\beta$, and $\mathbf{u}$ is optimal,

$$f(\boldsymbol{w}_t) - f(\mathbf{u}) \leq \exp\left(-{}^{t\lambda}\!/\!_{\beta}\right).$$

Smooth and strongly convex.

Gradient descent: $\boldsymbol{w}_0$ given; $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$.

If f is β-smooth and λ-strongly-convex, $\alpha_i = 1/\beta$, and $\mathbf{u}$ is optimal,

$$f(\boldsymbol{w}_t) - f(\mathbf{u}) \le \exp\left(-{}^{t\lambda}/_{\beta}\right).$$

Note. ${}^{\beta}/_{\lambda}$ is a **condition number**.

Example. Ridge regression $w \mapsto \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\lambda}{2}\|w\|^2$ is $(\lambda_{\max}(X^\top X) + \lambda)$-smooth and $(\lambda_{\min}(X^\top X) + \lambda)$-strongly-convex. Increasing λ brings condition number closer to 1.

Inexact gradients — descent directions.

We don't quite need $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \nabla f(\boldsymbol{w}_{i-1})$...

...can do $\boldsymbol{w}_i := \boldsymbol{w}_{i-1} - \alpha_i \boldsymbol{v}_i$ with $\boldsymbol{v}_i^\top \nabla(\boldsymbol{w}_{i-1}) > 0$;
in this case, $\boldsymbol{v}_i$ is a **descent direction**.

Can still prove rates.

Inexact gradients — stochastic gradients.

Can replace gradient **stochastic gradient** $\boldsymbol{v}_i$;
namely, $\mathbb{E}(\boldsymbol{v}_i) = \nabla f(\boldsymbol{w}_{i-1})$.

Example. With linear prediction, we can use

$$\boldsymbol{v}_i := y^{(i)}\mathbf{x}^{(i)}\ell'\left(\boldsymbol{w}_{i-1}^\top(y^{(i)}\mathbf{x}^{(i)})\right),$$

or a **minibatch**

$$\boldsymbol{v}_i := \frac{1}{|S_i|}\sum_{(\boldsymbol{x},y)\in S_i} y\boldsymbol{x}\ell'\left(\boldsymbol{w}_{i-1}^\top(y\boldsymbol{x})\right).$$

Rates. $1/\sqrt{t}$ (Lipschitz f, bounded domain), or $1/t$ (strong convexity).

Why? "Batch" gradient is expensive!

Other methods.

- Momentum/Acceleration (rate $1/t^2$ or $\exp(-t\sqrt{\lambda/\beta})$.
- Line searches.
- SVRG.
- ADMM.
- $\ldots$

Key topics.

- Logistic regression is glorious.
- Gradient descent hates bumps.
- Convex sets.
- Convex functions.
- ERM with convex losses of linear predictors.
- Gradient descent convergence rates.