Optimization for Machine Learning

DS3 Summer School

Martin Jaggi
EPFL

github.com/epfml/opt-summerschool

June 27, 2019

Logistics

- material based largely on EPFL course CS-439, with some new additions
- Additional contents available as a link on

  `github.com/epfml/opt-summerschool`

  - lecture notes
  - more exercises
  - more slide chapters

Outline

- Convexity, **Gradient Methods**, Constrained Optimization, Proximal algorithms, **Subgradient Methods**, **Stochastic Gradient Descent**, Coordinate Descent, Frank-Wolfe, **Accelerated Methods**, **Adaptive Methods**, Primal-dual context and certificates, Lagrange and Fenchel Duality, Second-Order methods including Quasi-Newton, Derivative-free optimization.
- Advanced Contents:
  - Parallel and Distributed Optimization Algorithms
  - **Non-Convex Optimization**: Convergence to Critical Points, Alternating minimization, **Neural network training**

Optimization

- General optimization problem (**unconstrained minimization**)

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{with} & \mathbf{x} \in \mathbb{R}^d \end{array}$$

  - candidate solutions, variables, parameters $\mathbf{x} \in \mathbb{R}^d$
  - objective function $f : \mathbb{R}^d \to \mathbb{R}$
  - typically: technical assumption: f is continuous and differentiable

Why? And How?

Optimization is everywhere

machine learning, big data, statistics, data analysis of all kinds, finance, logistics, planning, control theory, mathematics, search engines, simulations, and many other applications ...

- **Mathematical Modeling***:*
  - *defining & modeling the optimization problem*
- **Computational Optimization***:*
  - *running an (appropriate) optimization algorithm*

Optimization for Machine Learning

- **Mathematical Modeling**:
  - defining & and measuring the machine learning model
- **Computational Optimization**:
  - learning the model parameters
- Theory vs. practice:
  - libraries are available, algorithms treated as "black box" by most practitioners
  - **Not here:** we look inside the algorithms and try to understand why and how fast they work!

Optimization Algorithms

- Optimization at large scale: **simplicity** rules!
- Main approaches:
  - **Gradient Descent**
  - **Stochastic Gradient Descent** (SGD)
  - **Coordinate Descent**
- History:
  - 1847: Cauchy proposes gradient descent
  - 1950s: Linear Programs, soon followed by non-linear, SGD
  - 1980s: General optimization, convergence theory
  - 2005-today: Large scale optimization, convergence of SGD

Part 1

Gradient Descent

The Algorithm

Get near to a minimum $\mathbf{x}^\star$ / close to the optimal value $f(\mathbf{x}^\star)$?

(Assumptions: $f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, has a global minimum $\mathbf{x}^\star$)

Goal: Find $\mathbf{x} \in \mathbb{R}^d$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \varepsilon.$$

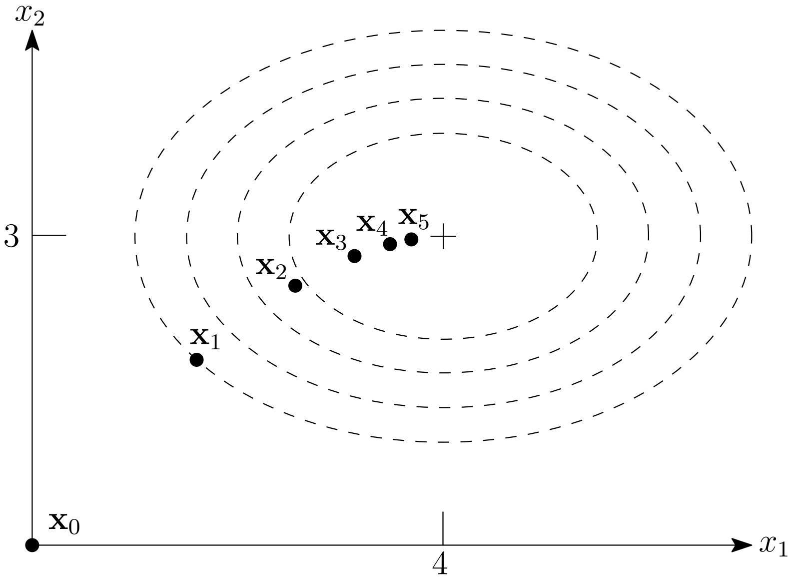
Note that there can be several minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

Iterative Algorithm:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),$$

for **timesteps** $t = 0, 1, \ldots,$ and **stepsize** $\gamma \geq 0$.

Example



x_2
3
$\mathbf{x}_5$
$\mathbf{x}_4$
$\mathbf{x}_3$
$\mathbf{x}_2$
$\mathbf{x}_1$
$\mathbf{x}_0$
4
x_1

Vanilla analysis

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$?

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$, and consider (using the definition of gradient descent)

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^\star).$$

- Apply $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ to rewrite

$$\begin{aligned} \mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) &= \frac{1}{2\gamma}\left(\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\right) \\ &= \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\right) \end{aligned}$$

- Sum this up over the iterations t:

$$\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) = \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_0-\mathbf{x}^\star\|^2 - \|\mathbf{x}_T-\mathbf{x}^\star\|^2\right)$$

Vanilla analysis, II

- Now we invoke convexity of f with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)$$

  giving

$$\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

  an upper bound for the average error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ over the steps

- last iterate is not necessarily the best one
- stepsize is crucial

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Assume that all gradients of f are bounded in norm.

- Equivalent to f being Lipschitz (Exercise 11).

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$. Choosing the stepsize

$$\gamma := \frac{R}{B\sqrt{T}},$$

gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}.$$

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps, II

Proof.

- Plug $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\mathbf{g}_t\| \leq B$ into Vanilla Analysis II:

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \leq \frac{\gamma}{2} B^2 T + \frac{1}{2\gamma} R^2.$$

- choose γ such that

$$q(\gamma) = \frac{\gamma}{2} B^2 T + \frac{R^2}{2\gamma}$$

  is minimized.
- Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{T}}$, and $q(R/(B\sqrt{T})) = RB\sqrt{T}$.
- Dividing by T, the result follows.

□

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps, III

$$T \geq \frac{R^2B^2}{\varepsilon^2} \quad \Rightarrow \quad \text{average error } \leq \frac{RB}{\sqrt{T}} \leq \varepsilon.$$

Advantages:

- dimension-independent!
- holds for both average, or best iterate

Smooth functions

"Not too curved"

Definition

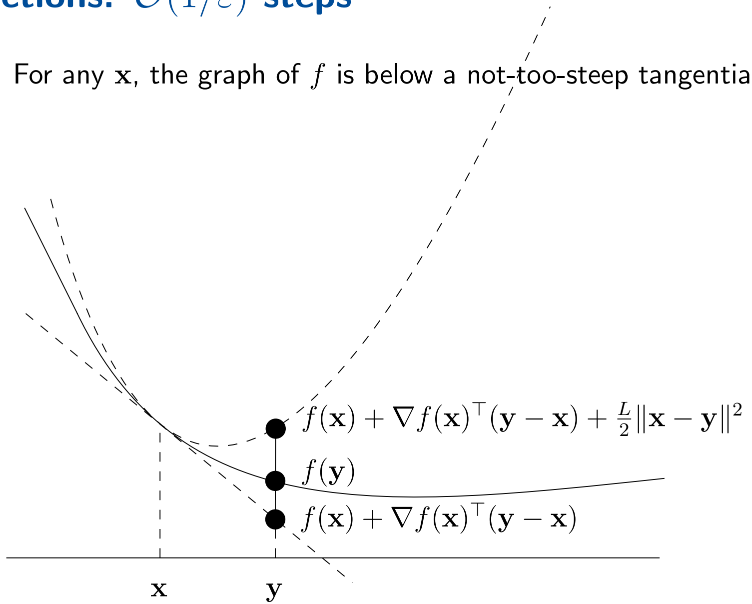
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. f is called smooth (with parameter $L \geq 0$) if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Definition does not require convexity (useful later)

Smooth functions: $\mathcal{O}(1/\varepsilon)$ steps

Smoothness: For any $\mathbf{x}$, the graph of f is below a not-too-steep tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



Smooth vs Lipschitz

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of f
- Smoothness $\Leftrightarrow$ Lipschitz continuity of ∇f (in the convex case).

Lemma

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.

(i) *f is smooth with parameter L.*

(ii) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Proof in lecture slides of L. Vandenberghe, `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf`.

Sufficient decrease

Lemma

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With

$$\gamma := \frac{1}{L},$$

gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

Note: More specifically, this already holds if f is smooth with parameter L over the line segment connecting $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Choosing stepsize

$$\gamma := \frac{1}{L},$$

gradient descent yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps II

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Vanilla Analysis II:

$$\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \le \frac{\gamma}{2}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

This time, we can bound the squared gradients by sufficient decrease:

$$\frac{1}{2L}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le \sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_T).$$

$\square$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps III

Putting it together with $\gamma = 1/L$:

$$\begin{aligned}\sum_{t=0}^{T-1} \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) &\leq \frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.\end{aligned}$$

Rewriting:

$$\sum_{t=1}^{T} \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

As last iterate is the best (sufficient decrease!):

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{T}\left(\sum_{t=1}^{T} \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right)\right) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps IV

$R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$

$$T \geq \frac{R^2 L}{2\varepsilon} \quad \Rightarrow \quad \text{error } \leq \frac{L}{2T} R^2 \leq \varepsilon.$$

- $50 \cdot R^2 L$ iterations for error 0.01 ...
- ... as opposed to $10,000 \cdot R^2 B^2$ in the Lipschitz case

Recap of gradient descent

Property of f	Learning Rate γ	Number of steps
$\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$	$\frac{R}{B\sqrt{T}}$	$\mathcal{O}(1/\varepsilon^2)$
f is L-smooth	$\frac{1}{L}$	$\mathcal{O}(1/\varepsilon)$

Part 2

Subgradient Descent

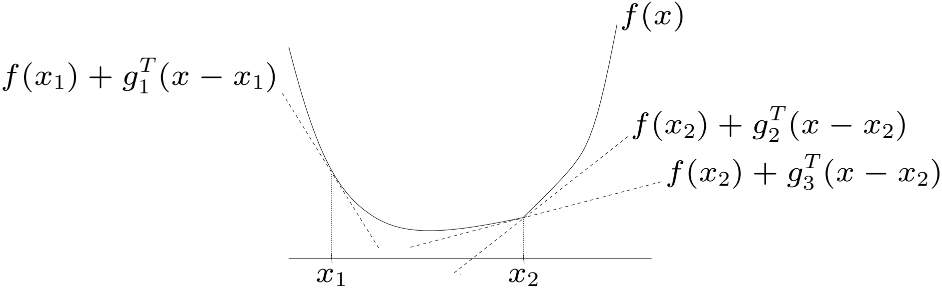
Subgradients

What if f is not differentiable?

Definition

$\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y}-\mathbf{x}) \qquad \text{for all } \mathbf{y} \in \mathbf{dom}(f)$$



$\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$ is the subdifferential, the set of subgradients of f at $\mathbf{x}$.

The subgradient descent algorithm

Subgradient descent: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

$$\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t$$

for **times** $t = 0, 1, \ldots$, and **stepsizes** $\gamma_t \geq 0$.

Stepsize can vary with time!

This is possible in (projected) gradient descent as well, but so far, we didn't need it.

Strongly convex functions

"Not too flat"

Straightforward generalization to the non-differentiable case:

Definition

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called strongly convex (with parameter μ) if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f),\ \forall \mathbf{g} \in \partial f(\mathbf{x}).$$

Strongly convex functions: characterization via "normal" convexity

Lemma (Exercise 28)

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ *be convex,* $\mathbf{dom}(f)$ *open,* $\mu \in \mathbb{R}_+, \mu > 0$. f *is strongly convex with parameter* μ *if and only if* $f_\mu : \mathbf{dom}(f) \to \mathbb{R}$ *defined by*

$$f_\mu(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|^2, \quad \mathbf{x} \in \mathbf{dom}(f)$$

is convex.

Tame strong convexity

For convergence, we assume that all subgradients $\mathbf{g}_t$ that we encounter during the algorithm are bounded in norm.

May be realistic if...

- we start close to optimality
- we run projected subgradient descent over a compact set X

May also fail!

- Over $\mathbb{R}^d$, strong convexity and bounded subgradients contradict each other! (Exercise 30).

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex with parameter $\mu > 0$ and let $\mathbf{x}^\star$ be the unique global minimum of f. With decreasing step size

$$\gamma_t := \frac{2}{\mu(t+1)}, \quad t > 0,$$

subgradient descent yields

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \le \frac{2B^2}{\mu(T+1)},$$

where $B = \max_{t=1}^{T} \|\mathbf{g}_t\|$. $\uparrow$

convex combination of iterates

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps II

Proof.

Vanilla analysis ($\mathbf{g}_t \in \partial f(\mathbf{x}_t)$):

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma_t}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma_t}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right).$$

Lower bound from strong convexity:

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Putting it together (with $\|\mathbf{g}_t\|^2 \leq B^2$):

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t^{-1}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2.$$

Summing over $t = 1, \ldots, T$: we used to have telescoping ($\gamma_t = \gamma, \mu = 0$)...

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps III

Proof.

So far we have:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - \frac{\gamma_t^{-1}}{2} \left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2.$$

To get telescoping, we would need $\gamma_t^{-1} = \gamma_{t+1}^{-1} - \mu$.

Works with $\gamma_t^{-1} = \mu(1+t)$, but not $\gamma_t^{-1} = \mu(1+t)/2$ (the choice here).

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps IV

Proof.

So far we have:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - \frac{\gamma_t^{-1}}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2.$$

Plug in $\gamma_t^{-1} = \mu(1+t)/2$ and multiply with t on both sides:

$$\begin{aligned} t\cdot\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) &\le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big) \\ &\le \frac{B^2}{\mu} + \frac{\mu}{4}\Big(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big). \end{aligned}$$

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps V

Proof.

We have

$$t \cdot \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big)$$
$$\le \frac{B^2}{\mu} + \frac{\mu}{4}\Big(t(t-1)\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|^2 - (t+1)t\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|^2\Big).$$

Now we get telescoping...

$$\sum_{t=1}^{T} t \cdot \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \le \frac{TB^2}{\mu} + \frac{\mu}{4}\Big(0 - T(T+1)\left\|\mathbf{x}_{T+1} - \mathbf{x}^\star\right\|^2\Big) \le \frac{TB^2}{\mu}.$$

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps VI

Proof.

Almost done:

$$\underline{\sum_{t=1}^{T} t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right)} \leq \frac{TB^2}{\mu} + \frac{\mu}{4}\left(0 - T(T+1)\left\|\mathbf{x}_{T+1} - \mathbf{x}^\star\right\|^2\right) \leq \frac{TB^2}{\mu}.$$

Since

$$\frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1,$$

Jensen's inequality yields

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2}{T(T+1)} \underline{\sum_{t=1}^{T} t \cdot \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right)}.$$

Tame strong convexity: Discussion

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2B^2}{\mu(T+1)},$$

Weighted average of iterates achieves the bound (later iterates have more weight)

Bound is independent of initial distance $\|\mathbf{x}_0 - \mathbf{x}^\star\|\ldots$

…but not really: B typically depends on $\|\mathbf{x}_0 - \mathbf{x}^\star\|$ (for example, $B = \mathcal{O}(\|\mathbf{x}_0 - \mathbf{x}^\star\|)$ for quadratic functions)

Recall: we can only hope that B is small (can be checked while running the algorithm)

What if we don't know the parameter μ of strong convexity?

$\rightarrow$ **Bad luck!** In practice, try some μ's, pick best solution obtained

Part 3

Stochastic Gradient Descent

Stochastic gradient descent

Many objective functions are sum structured:

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}).$$

Example: f_i is the cost function of the i-th observation, taken from a training set of n observation.

Evaluating $\nabla f(\mathbf{x})$ of a sum-structured function is expensive (sum of n gradients).

Stochastic gradient descent: the algorithm

choose $\mathbf{x}_0 \in \mathbb{R}^d$.

sample $i \in [n]$ uniformly at random

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t).$$

for **times** $t = 0, 1, \ldots,$ and **stepsizes** $\gamma_t \geq 0$.

Only update with the gradient of f_i instead of the full gradient!

Iteration is n times cheaper than in full gradient descent.

The vector $\mathbf{g}_t := \nabla f_i(\mathbf{x}_t)$ is called a stochastic gradient.

$\mathbf{g}_t$ is a vector of d random variables, but we will also simply call this a random variable.

Unbiasedness

Can't use convexity

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$$

on top of the vanilla analysis, as this may hold or not hold, depending on how the stochastic gradient $\mathbf{g}_t$ turns out.

We will show (and exploit): the inequality holds in expectation.

Fot this, we use that by definition, $\mathbf{g}_t$ is an **unbiased estimate** of $\nabla f(\mathbf{x}_t)$:

$$\mathbb{E}\left[\mathbf{g}_t \middle| \mathbf{x}_t = \mathbf{x}\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

The inequality $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ holds in expectation

For any fixed $\mathbf{x}$, linearity of conditional expectations yields

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x} - \mathbf{x}^\star)\big|\mathbf{x}_t = \mathbf{x}\big] = \mathbb{E}\big[\mathbf{g}_t\big|\mathbf{x}_t = \mathbf{x}\big]^\top(\mathbf{x} - \mathbf{x}^\star) = \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star).$$

So $\downarrow$ convexity

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \geq \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big].$$

The first equality is concerning? The event $\{\mathbf{x}_t = \mathbf{x}\}$ can occur only for $\mathbf{x}$ in some finite set X ($\mathbf{x}_t$ is determined by the choices of indices in all iterations so far).
By the so called Partition Theorem (Exercise 32):

$$\begin{aligned}\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] &= \sum_{\mathbf{x}\in X} \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x} - \mathbf{x}^\star)\big|\mathbf{x}_t = \mathbf{x}\big] \operatorname{prob}(\mathbf{x}_t = \mathbf{x}) \\ &= \sum_{\mathbf{x}\in X} \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star) \operatorname{prob}(\mathbf{x}_t = \mathbf{x}) = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big].\end{aligned}$$

Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be convex and differentiable,* $\mathbf{x}^\star$ *a global minimum; furthermore, suppose that* $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$*, and that* $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$ *for all* t*. Choosing the constant stepsize*

$$\gamma := \frac{R}{B\sqrt{T}}$$

stochastic gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[f(\mathbf{x}_t)\big] - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}.$$

Same procedure as every time. . . except

- we assume bounded stochastic gradients in expectation;
- error bound holds in expectation.

Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps II

Proof.

Vanilla analysis (this time, $\mathbf{g}_t$ is the stochastic gradient):

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Taking expectations and using "convexity in expectation":

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] \leq \sum_{t=0}^{T-1} \mathbb{E}\big[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)\big] \quad &\leq \quad \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\mathbf{g}_t\|^2\big] + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \\
&\leq \quad \frac{\gamma}{2} B^2 T + \frac{1}{2\gamma} R^2.
\end{aligned}$$

Result follows as every time (optimize γ) ... $\square$

Recap of GD and SGD

Gradient Descent:

Property of f	Learning Rate γ	Number of steps
$\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$	$\frac{R}{B\sqrt{T}}$	$\mathcal{O}(1/\varepsilon^2)$

Stochastic Gradient Descent:

Property of f	Learning Rate γ	Number of steps
$\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \leq B^2$ for all t	$\frac{R}{B\sqrt{T}}$	$\mathcal{O}(1/\varepsilon^2)$

Convergence rate comparison: SGD vs GD

Classic GD: For vanilla analysis, we assumed that $\|\nabla f(\mathbf{x})\|^2 \leq B_{\mathsf{GD}}^2$ for all $\mathbf{x} \in \mathbb{R}^d$, where B_{GD} was a constant. So for sum-objective:

$$\left\|\frac{1}{n}\sum_i \nabla f_i(\mathbf{x})\right\|^2 \leq B_{\mathsf{GD}}^2 \qquad \forall \mathbf{x}$$

SGD: Assuming same for the expected squared norms of our stochastic gradients, now called B_{SGD}^2.

$$\frac{1}{n}\sum_i \left\|\nabla f_i(\mathbf{x})\right\|^2 \leq B_{\mathsf{SGD}}^2 \qquad \forall \mathbf{x}$$

So by Jensen's inequality for $\|.\|^2$

- $B_{\mathsf{GD}}^2 \approx \left\|\frac{1}{n}\sum_i \nabla f_i(\mathbf{x})\right\|^2 \leq \frac{1}{n}\sum_i \left\|\nabla f_i(\mathbf{x})\right\|^2 \approx B_{\mathsf{SGD}}^2$
- B_{GD}^2 can be smaller than B_{SGD}^2, but often comparable.
  Very similar if larger mini-batches are used.

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$; let $\mathbf{x}^\star$ be the unique global minimum of f. With decreasing step size

$$\gamma_t := \frac{2}{\mu(t+1)}$$

stochastic gradient descent yields

$$\mathbb{E}\Big[f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\Big) - f(\mathbf{x}^\star)\Big] \le \frac{2B^2}{\mu(T+1)},$$

where $B^2 := \max_{t=1}^{T} \mathbb{E}\big[\, \|\mathbf{g}_t\|^2 \,\big]$.

Almost same result as for subgradient descent, but in expectation.

Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps II

Proof.

Take expectations over vanilla analysis, before summing up (with varying stepsize γ_t):

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] = \frac{\gamma_t}{2}\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] + \frac{1}{2\gamma_t}\left(\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}^\star\|^2\big] - \mathbb{E}\big[\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\big]\right).$$

"Strong convexity in expectation":

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \geq \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] + \frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}^\star\|^2\big]$$

Putting it together (with $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$):

$$\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)] \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}^\star\|^2\big] - \frac{\gamma_t^{-1}}{2}\mathbb{E}\big[\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\big].$$

Proof continues as for subgradient descent, this time with expectations. $\square$

Mini-batch SGD

Instead of using a single element f_i, use an average of several of them:

$$\tilde{\mathbf{g}}_t := \frac{1}{m} \sum_{j=1}^{m} \mathbf{g}_t^j .$$

Extreme cases:

$m = 1 \Leftrightarrow$ SGD as originally defined

$m = n \Leftrightarrow$ full gradient descent

Benefit: Gradient computation can be naively parallelized

Mini-batch SGD

Variance Intuition: Taking an average of many independent random variables reduces the variance. So for larger size of the mini-batch m, $\tilde{\mathbf{g}}_t$ will be closer to the true gradient, in expectation:

$$
\begin{aligned}
\mathbb{E}\Big[\Big\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\Big\|^2\Big] =& \mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_t^j - \nabla f(\mathbf{x}_t)\Big\|^2\Big] \\
=& \frac{1}{m}\mathbb{E}\big[\|\mathbf{g}_t^1 - \nabla f(\mathbf{x}_t)\|^2\big] \\
=& \frac{1}{m}\mathbb{E}\big[\|\mathbf{g}_t^1\|^2\big] - \frac{1}{m}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{B^2}{m} \ .
\end{aligned}
$$

Using a modification of the SGD analysis, can use this quantity to relate convergence rate to the rate of full gradient descent.

Stochastic Subgradient Descent

For problems which are not necessarily differentiable, we modify SGD to use a subgradient of f_i in each iteration. The update of **stochastic subgradient descent** is given by

sample $i \in [n]$ uniformly at random

let $\mathbf{g}_t \in \partial f_i(\mathbf{x}_t)$

$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t.$

In other words, we are using an unbiased estimate of a subgradient at each step, $\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big] \in \partial f(\mathbf{x}_t)$.

Convergence in $\mathcal{O}(1/\varepsilon^2)$, by using the subgradient property at the beginning of the proof, where convexity was applied.

Constrained optimization

For constrained optimization, our theorem for the SGD convergence in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to constrained problems as well.

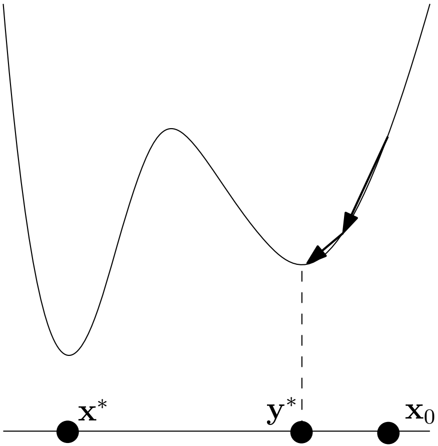
After every step of SGD, projection back to X is applied as usual. The resulting algorithm is called projected SGD.

Part 4

Non-convex Optimization

Gradient Descent in the nonconvex world

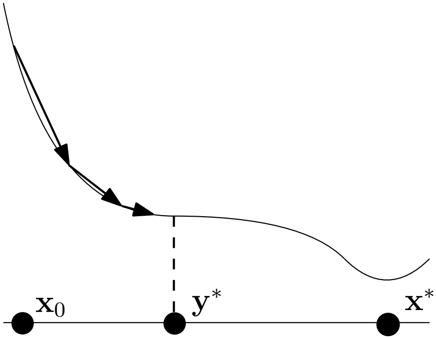
- may get stuck in a local minimum and miss the global minimum;

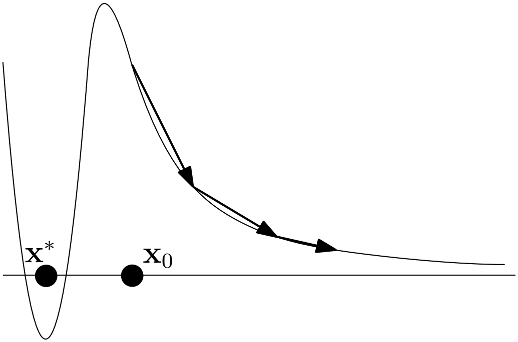


Gradient Descent in the nonconvex world II

Even if there is a unique local minimum (equal to the global minimum), we

- may get stuck in a saddle point;
- run off to infinity;
- possibly encounter other bad behaviors.





Gradient Descent in the nonconvex world III

Often, we observe good behavior in practice.

Theoretical explanations mostly missing.

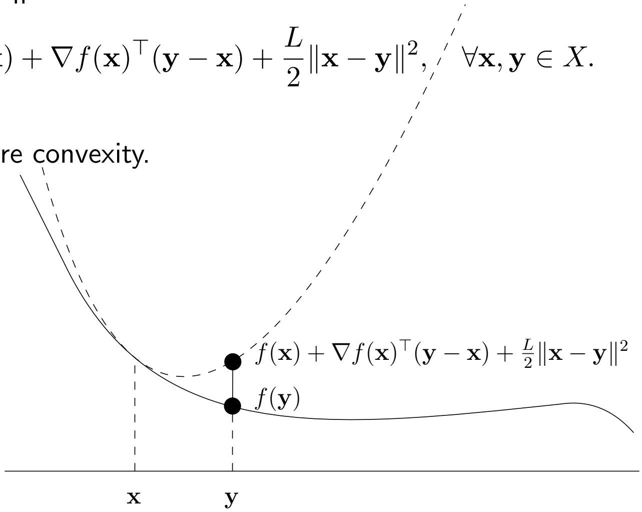
This lecture: under favorable conditions, we sometimes can say something useful about the behavior of gradient descent, even on nonconvex functions.

Smooth (but not necessarily convex) functions

Recall: A differentiable $f : \mathbf{dom}(f) \to \mathbb{R}$ is smooth with parameter $L \in \mathbb{R}_+$ over a convex set $X \subseteq \mathbf{dom}(f)$ if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X. \tag{1}$$

Definition does not require convexity.

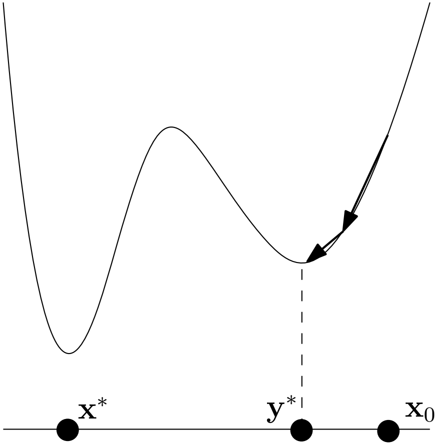


Gradient descent on smooth functions

Will prove: $\|\nabla f(\mathbf{x}_t)\|^2 \to 0$ for $t \to \infty$...

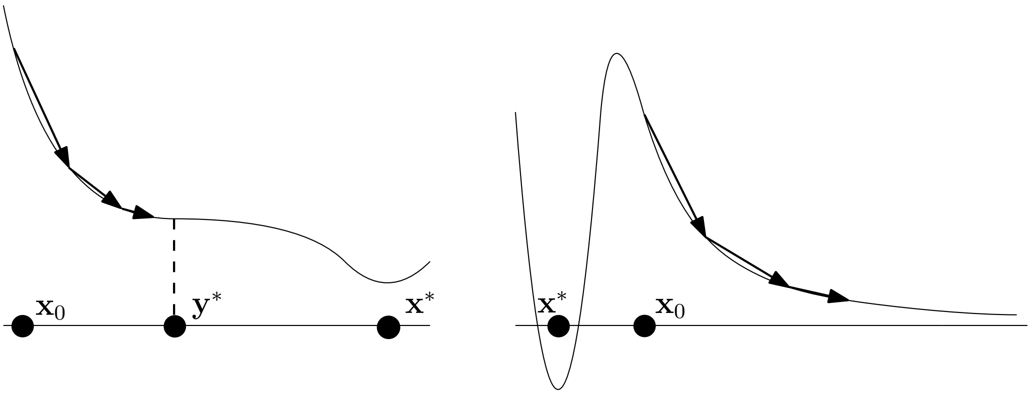
...at the same rate as $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \to 0$ in the convex case.

$f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ itself may not converge to 0 in the nonconvex case:



What does $\|\nabla f(\mathbf{x}_t)\|^2 \to 0$ mean?

It may or may not mean that we converge to a **critical point** ($\nabla f(\mathbf{y}^\star) = \mathbf{0}$)



Gradient descent on smooth (not necessarily convex) functions

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L according to Definition 2. Choosing stepsize

$$\gamma := \frac{1}{L},$$

gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big), \quad T > 0.$$

In particular, $\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{2L}{T}\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big)$ *for some* $t \in \{0, \ldots, T-1\}$.

And also, $\lim_{t\to\infty} \|\nabla f(\mathbf{x}_t)\|^2 = 0$ *(Exercise 34).*

Gradient descent on smooth (not necessarily convex) functions II

Proof.

Sufficient decrease (see above) does not require convexity:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

Rewriting:

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2L\big(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})\big).$$

Telescoping sum:

$$\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq 2L\big(f(\mathbf{x}_0) - f(\mathbf{x}_T)\big) \leq 2L\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big).$$

The statement follows (divide by T). $\square$

Part 5

Improvements to SGD

Momentum

Idea:
Use **momentum** from "movement" so far

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t) + \nu \big[\mathbf{x}_t - \mathbf{x}_{t-1} \big]$$

$\nu > 0$ is called the momentum parameter

Adagrad

Adagrad is an adaptive variant of SGD

pick a stochastic gradient $\mathbf{g}_t$

update $[G_t]_i := \sum_{s=0}^{t} ([\mathbf{g}_s]_i)^2$ for each feature i

$[\mathbf{x}_{t+1}]_i := [\mathbf{x}_t]_i - \frac{\gamma}{\sqrt{[G_t]_i}} [\mathbf{g}_t]_i$ for each feature i

(standard choice of $\mathbf{g}_t := \nabla f_j(\mathbf{x}_t)$ for sum-structured objective functions $f = \sum_j f_j$)

- chooses an adaptive, coordinate-wise learning rate
- strong performance in practice
- Variants: Adadelta, Adam, RMSprop

SignSGD

Only use the sign (one bit) of each gradient entry:
SignSGD is a communication efficient variant of SGD.

pick a stochastic gradient $\mathbf{g}_t$

$$[\mathbf{x}_{t+1}]_i := [\mathbf{x}_t]_i - \gamma \, sign([\mathbf{g}_t]_i) \quad \text{for each feature } i$$

(standard choice of $\mathbf{g}_t := \nabla f_j(\mathbf{x}_t)$ for sum-structured objective functions $f = \sum_j f_j$)

- communication efficient for distributed training
- convergence issues

Part 6 / Afternoon

Try it yourself!

Convergence of the discussed algorithms in action,

- for training deep networks, and
- other optimization problems related to deep learning (e.g. style transfer, adversarial examples)

These slides, and additional materials:

github.com/epfml/opt-summerschool