Deep Learning

Lecture 04

Optimization

Aurelien Lucchi

Fall 2024

Section 1

Objectives

Loss Functions: Examples

Given target/ground-truth $\mathbf{y}$ and prediction ν.

Squared loss

$$\ell_{\mathbf{y}}(\boldsymbol{\nu}) = \frac{1}{2}\|\mathbf{y} - \boldsymbol{\nu}\|^2$$

Zero-one loss

$$\ell_y(\nu) = \begin{cases} 0, & \text{if } \nu = y \\ 1, & \text{else} \end{cases}$$

Log-loss (multiclass)

$$\ell_y(\nu) = -\log \nu_y$$

Soft target cross-entropy ($\mathbf{y} \in [0;1]^m$)

$$\ell_{\mathbf{y}}(\boldsymbol{\nu}) = -\sum_{j=1}^{m} y_j \log \nu_j \geq -\sum_{j=1}^{m} y_j \log y_j =: H(\mathbf{y})$$

Expected and Empirical Risk

Loss functions are defined on single instances.

Taking expectations over loss functions: **risk function**

Expected risk: $(\xi, \mathbf{y}) \in \mathbb{P}$ (unknown data generating distribution)

$$\mathcal{R}(\mathbf{x}) = \mathbf{E}_{\mathbb{P}}[\ell(F(\mathbf{x}; \xi)]$$

Ideally $\mathbf{x}^* = \mathrm{argmin}_{\mathbf{x}} \mathcal{R}(\mathbf{x})$.

Empirical risk: in practice only access to a sample $\mathcal{S} = \{(\mathbf{x}^t, \mathbf{y}^t)\}$

$$\mathcal{R}_{\mathcal{S}}(\mathbf{x}) = \frac{1}{s} \sum_{t=1}^{s} \ell_{\mathbf{y}^t}(F(\mathbf{x}^t; \mathbf{x}))$$

F: $\boldsymbol{\theta}$-parameterized model (e.g. DNN).

Minimizing empirical risk as a proxy for expected risk: **empirical risk minimization**.

Section 2

Optimality

Function class

Goal: Given $f : \mathbb{R}^d \to \mathbb{R}$, find $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} f(\mathbf{x})$.

Function class: f typically exhibit certain properties:

- C^0 = class of all continuous functions
- C^1 = all differentiable functions whose derivative is continuous
- Next, we will review the class of convex functions

Convex set

Definition 1 (Convex set)

A set S is convex if $\forall \mathbf{x}, \mathbf{y} \in S$, $\lambda \in [0, 1]$,

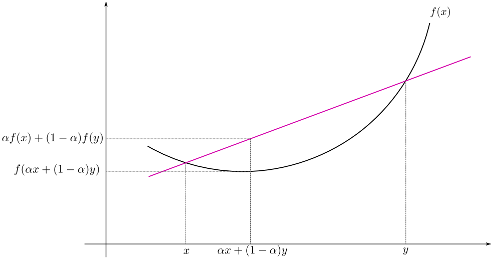
$$\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} \in S$$

Convex Functions

Definition 2 (Convex function)

A function $f : S \to \mathbb{R}$ is convex if its domain S is a convex set and if for any two points $\mathbf{x}, \mathbf{y} \in S$, the following property holds:

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



Local vs Global minimum

Necessary and Sufficient Conditions

Necessary condition: used to determine the points where a function may achieve a critical point.
$\rightsquigarrow$ "necessary" because for a point to be an extremum (maximum or minimum), this condition must hold
$\rightsquigarrow$ However, it is not sufficient by itself to guarantee an extremum.

Sufficient condition: guarantees that the point is indeed a local maximum or minimum.

Necessary and Sufficient Conditions

First-order necessary conditions

Assume $f \in C^1(\mathbb{R}^d)$. If $\mathbf{x}^* \in \mathbb{R}^d$ is a local minimizer of f, then $\nabla f(\mathbf{x}^*) = 0$.

Second-order necessary conditions

Assume $f \in C^2(\mathbb{R}^d)$. Then $\mathbf{x}^*$ is a local minimizer of f implies $\nabla^2 f(\mathbf{x}^*)$ is positive semidefinite, i.e. $\mathbf{s}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{s} \geq 0$ for all $\mathbf{s} \in \mathbb{R}^d$.

Second-order sufficient conditions

Assume $f \in C^2(\mathbb{R}^d)$. Then $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite (i.e. $\mathbf{s}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{s} \geq 0$ for all $\mathbf{s} \in \mathbb{R}^d$) implies that $\mathbf{x}^*$ is a local minimizer of f.

Section 3

Gradient Descent

Gradient Descent: Iterate Sequence

Remark: Note the change of notation to avoid too much clutter.

Goal: Minimize a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$.

Gradient Descent: Consider the evolution of the iterates

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$$

starting from some initial $\mathbf{x}_0$.

Fixed step size $\eta > 0$.

Step Size

Important question in gradient descent: choice of the step size.

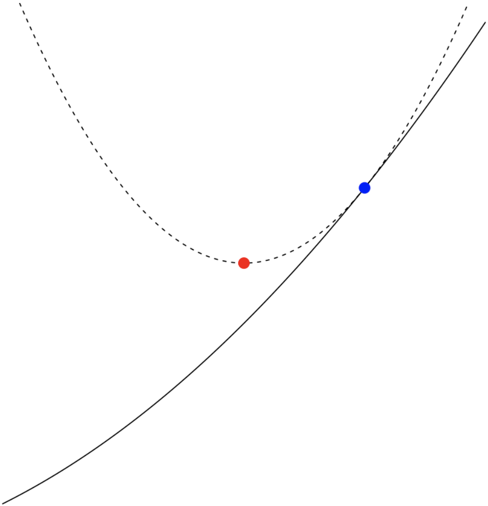
- Is η small enough such that discretization approximates gradient flow?
- Is η small enough to ensure convergence towards a minimizer?

Larger step sizes are usually preferred from the standpoint of computational complexity and performance.

Quadratic Model

Taylor expansion: Consider local approximation of f:

$$f(\mathbf{x}+\Delta\mathbf{x}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^\top \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}$$



Quadratic Model

Taylor expansion: Consider local approximation of f:

$$f(\mathbf{x}+\Delta\mathbf{x}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^\top \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}$$

Minimizer over $\Delta\mathbf{x}$ is:

$$\Delta\mathbf{x} = -[\nabla^2 f(\mathbf{x})]^{-1} \nabla f(\mathbf{x})$$

This is **Newton's method**. It requires computing and inverting the Hessian (expensive in high dimensions!).

Preconditioning Can we choose a different matrix $\mathbf{B}$ that is close to the Hessian (i.e. $\|\nabla^2 f(\mathbf{x}) - \mathbf{B}\|_2 \leq \epsilon$)?

Quadratic Model, Pre-Conditioner

Recovering Gradient Descent: If $\nabla^2 f(\mathbf{x}) = \mathbf{I}$, then

$$f(\mathbf{x} + \Delta\mathbf{x}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \Delta\mathbf{x} + \frac{1}{2\eta}\|\Delta\mathbf{x}\|^2$$

Minimizer is just $\Delta\mathbf{x} = -\eta\nabla f(\mathbf{x})$.

Diagonal approximation

$$\mathbf{B} = \mathsf{Diag}(\nabla^2 f(\mathbf{x}))$$

Cheaper to compute and easy to invert. But it might not always be a good approximation.

Minimizing a Convex Quadratic

Goal: Analyse the gradient dynamics and understand its convergence.

Start with convex quadratic objective

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\,\mathbf{x} - \mathbf{q}^\top \mathbf{x}, \quad \mathbf{Q} \text{ positive definite}$$

Diagonalize $\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ with orthogonal $\mathbf{U}$.

Change of basis $\mathbf{x} \leftarrow \mathbf{U}^\top \mathbf{x}$, $\mathbf{q} \leftarrow \mathbf{U}^\top \mathbf{q}$.

Effectively minimize the **separable** problem

$$f(\mathbf{x}) = \sum_{i=1}^{n} g_i(x_i), \quad g_i(z) = \frac{\lambda_i}{2} z^2 - q_i z, \quad \lambda_i > 0\,.$$

Minimizing a Convex Quadratic

Look at generic $g := g_i : \mathbb{R} \to \mathbb{R}$, $g(z) = \frac{\lambda}{2} z^2 - qz$

Derivative:

$$g'(z) = \lambda z - q$$

Minimizer:

$$z^* = q/\lambda, \quad g(z^*) = \frac{q^2}{2\lambda} - \frac{q^2}{\lambda} = -\frac{q^2}{2\lambda}$$

Simplify analysis: shift g

$$g(z) \leftarrow g(z) - g(z^*) = \frac{\lambda}{2}\left(z^2 - \frac{2q}{\lambda} z + \frac{q^2}{\lambda^2}\right) = \frac{1}{2\lambda}\left(\lambda z - q\right)^2$$

Convex Quadratic: Gradient Step

A gradient step results in

$$\begin{aligned} g(z - \eta(\lambda z - q)) &= \frac{1}{2\lambda}\left((1 - \lambda\eta)(\lambda z - q)\right)^2 \\ &= (1 - \lambda\eta)^2 g(z) \end{aligned}$$

Requirement: $\eta < 2/\lambda$ for the objective decrease by a constant factor < 1 in every step.

Exponentially fast convergence to minimum.

Convex Quadratic: Optimal Step Size

Coming back to multi-dimensional quadratic: The step size condition becomes

$$\eta \leq \frac{2}{\lambda_{max}(\mathbf{Q})}$$

Optimal step size:

$$\begin{aligned} \eta^* &= \min_{\eta} \max_{i} (1 - \eta\lambda_i)^2 \\ &= \min_{\eta} \max\{\eta\lambda_{\max} - 1, 1 - \eta\lambda_{min}\} \end{aligned}$$

attained at

$$\eta\lambda_{max} - 1 \overset{!}{=} 1 - \eta\lambda_{min}$$

$$\iff \eta^* = \frac{2}{\lambda_{max} + \lambda_{min}}$$

Convex Quadratic: Optimal Rates

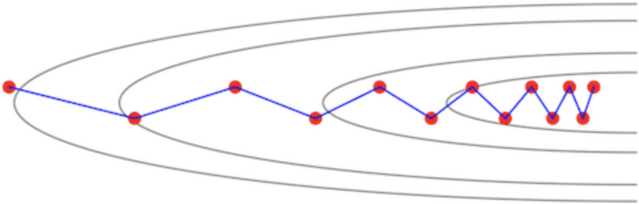
Slowest rate (in direction of eigenvector with smallest eigenvalue)

$$\rho = (1 - \lambda_{min}\eta^*)^2 = \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}\right)^2$$

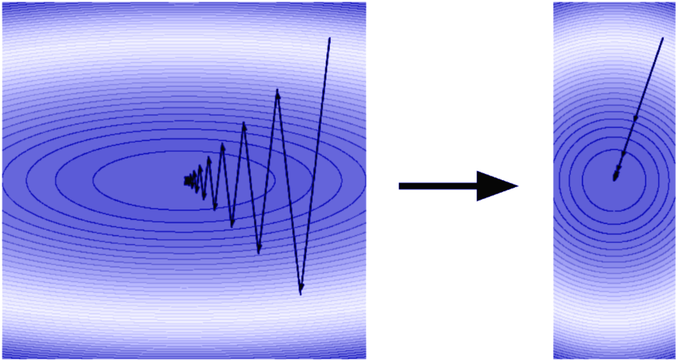
$$\leq (1 - \kappa)^2, \quad \kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

κ: condition number of $\mathbf{Q}$.

Example



Preconditioning



Smoothness and Convexity

L-smooth function (L-Lipschitz-continuous gradient):

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x} + \Delta\mathbf{x})\| \leq L\|\Delta\mathbf{x}\|$$

μ-strongly convex function

$$f(\mathbf{x} + \Delta\mathbf{x}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \Delta\mathbf{x} + \frac{\mu}{2}\|\Delta\mathbf{x}\|^2$$

Convex if $\mu = 0$.

Smoothness and Convexity

If f is twice differentiable then smoothness can be restated as

$$\boxed{f(\mathbf{x} + \Delta\mathbf{x}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \Delta\mathbf{x} + \frac{L}{2}\|\Delta\mathbf{x}\|^2}$$

The Hessian of a (twice differentiable) smooth and strongly convex function is thus sandwiched

$$\boxed{\mu\mathbf{I} \preceq \nabla^2 f \preceq L\mathbf{I}}$$

This connects back to the quadratic case, where $\lambda_{\min} = \mu$ and $\lambda_{max} = L$.

Convergence: Strongly Convex Case

Theorem 3

For a μ-strongly convex, L-smooth function f, the gradient descent iterates $\mathbf{x}_k$ with step size $0 < \eta \leq 1/L$ converge to the unique minimizer $\mathbf{x}^*$ at rate

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq (1 - \eta\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Corollary: Using L-smoothness, we also have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \frac{L}{2}(1 - \eta\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Lemma

The proof makes use of the following lemma:

Lemma 4

f is differentiable and L-smooth. Then

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2$$

Proof.

$$\begin{aligned} f(\mathbf{x}^*) - f(\mathbf{x}) &\leq f\left(\mathbf{x} - \tfrac{1}{L}\nabla f(\mathbf{x})\right) - f(\mathbf{x}) \\ &\leq f(\mathbf{x}) - \tfrac{1}{L}\|\nabla f(\mathbf{x})\|^2 + \frac{L}{2}\|\tfrac{1}{L}\nabla f(\mathbf{x})\|^2 - f(\mathbf{x}) \\ &= -\frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \end{aligned}$$

□

Convergence: Strongly Convex Case

Proof: Use $f(\mathbf{x}^*) - f(\mathbf{x}) \geq \nabla f(\mathbf{x}) \cdot (\mathbf{x}^* - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}\|^2$.

Convergence: Strongly Convex Case

Proof.
By induction using

$$
\begin{aligned}
&\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \\
&= \|\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k) - \mathbf{x}^*\|^2 \\
&= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \nabla f(\mathbf{x}_k) \cdot (\mathbf{x}_k - \mathbf{x}^*) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \\
&\overset{\mu}{\leq} (1 - \eta\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \\
&\overset{L}{\leq} (1 - \eta\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + 2L\eta^2(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \\
&= (1 - \eta\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta(1 - \eta L)(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \\
&\leq (1 - \eta\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2
\end{aligned}
$$

□

Convergence: Convex (and Smooth) Functions

The convexity condition alone (with smoothness) can only guarantee a slower convergence.

Theorem 5

Let f be convex, differentiable and l-smooth. Then with step size $\eta \leq \frac{1}{L}$, the suboptimality along the gradient descent iterates decays like

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{2\eta k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

where $\mathbf{x}^*$ is a minimizer of f.

ϵ-Stationarity

Non-convex case: local convergence measured in terms of a vanishing gradient norm.

$$\boxed{\|\nabla f(\mathbf{x})\| \leq \epsilon} \qquad (\epsilon\text{-stationarity})$$

Non-Convex Case

Theorem 6

f: differentiable, L-smooth, not necessarily convex with minimum f^*. The gradient descent iterates with step size $\eta \leq 1/L$ satisfy

$$\min_{i=0}^{k} \|\nabla f(\mathbf{x}_i)\|^2 \leq \frac{2(f(\mathbf{x}_0) - f^*)}{\eta(k+1)}$$

Note that in the convex case, the decay of the iterate suboptimality also implies that the squared gradient norm vanishes at the same rate. Here the gradient norm vanishes at rate $1/k$ without making any statement about the suboptimality of the iterates.

Descent Lemma

Lemma 7 (Descent lemma)

Consider a gradient descent step $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$ *on a* L*-smooth function* f*. For* $\eta \leq 1/L$ *this yields the following function decrease:*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\eta}{2} \|\nabla f(\mathbf{x}_k)\|^2. \tag{1}$$

Non-Convex Case: Proof Theorem 6

1. For any $\eta \leq 1/L$ by the Lemma above

$$\frac{\eta}{2}\|\nabla f(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$$

2. Summing over iterates we get a telescoping sum

$$\frac{\eta}{2}\sum_{i=0}^{k}\|\nabla f(\mathbf{x}_i)\|^2 \leq f(\mathbf{x}_0) - f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_0) - f(\mathbf{x}^*)$$

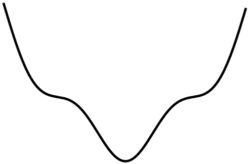
3. We can lower bound the sum by the smallest summand

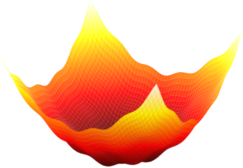
$$\sum_{i=0}^{k}\|\nabla f(\mathbf{x}_i)\|^2 \geq (k+1)\min_{i=0}^{k}\|\nabla f(\mathbf{x}_i)\|^2$$

Polyak-Łojasiewicz Condition

A generalization of strong convexity without the convexity :)

$$\boxed{\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^*),\ \forall \mathbf{x},\ f^* = \min f(\mathbf{x})}$$





PL-condition ensures fast local convergence: squared gradient norm will not vanish faster than the suboptimality (see next theorem)!

Exercise: show that a μ-strongly convex f fulfills the PL condition.

Convergence Theorem with PL Condition

Theorem 8

Let f be differentiable and L-smooth, not necessarily convex with minimum f^* and fulfilling the PL condition with $\mu > 0$. The gradient descent iterates with step size $\eta \leq 1/L$ satisfy

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(\mathbf{x}_0) - f^*)$$

PL Convergence Theorem: Proof

1. Descent lemma for L-smooth functions

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2$$

2. PL condition

$$-\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \leq -\frac{\mu}{L}(f(\mathbf{x}_k) - f^*)$$

3. Subtracting f^* on both sides

$$f(\mathbf{x}_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}_k) - f^*)$$

The claim follows by induction.

Challenges: Local Minima

[cf. DL, Section 8.2.3]

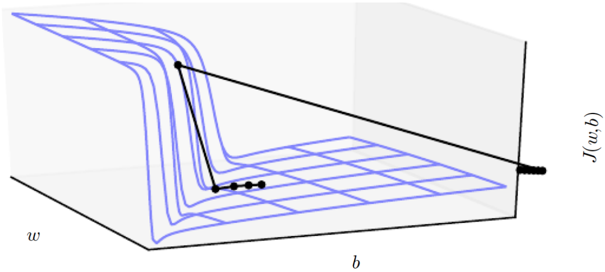
If neural network risk functions have many local minima and/or saddle points, then gradient descent would get stuck.

1. Are local minima a practical issue? Not always: Gori & Tesi, 1992
2. Do local minima even exist? Sometimes not (auto-encoder): Baldi & Hornik, 1989
3. Are local minima typically worse? Often not (large networks): e.g. Choromanska et al, 2015
4. Which local minima generalize well (wide vs. sharp.isolated): ongoing discussion
5. Can we understand the learning dynamics? Deep linear case has similarities with non-linear case, e.g. Saxe et al., 2013

Challenges: Local Minima

[cf. DL, Section 8.2.3]

Models with multiplication of many weights (depth, recurrence): sharp non-linearities



Motivates gradient clipping heuristics.

Section 4

Stochastic Gradient Descent

Motivation

- Computational complexity of gradient descent: **linear** in number of samples. Prohibitive for large data sets
- Approximate empirical average over all training instances by an empirical average on a smaller sample?
- Fixed subsampling: loss of statistical efficiency
- **Key Idea:** subsample at every update step when computing update directions!

Stochastic Gradient Descent

Assume additive form (e.g. empirical risk)

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}), \quad \text{s.t.} \quad \nabla f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x})$$

SGD

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f_{I(k)}(\mathbf{x}_k), \quad I(k) \sim \text{Uniform}\{1, \ldots, n\}}$$

$\eta_k > 0$ is a step size schedule (e.g. decaying with k).

SGD is a classical method going back to the 1950's.

Bias and Variance

Update direction of SGD is **unbiased**

$$\mathbf{E}[\nabla f_I(\mathbf{x})] = \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$$

Stochasticity can be quantified by variance function

$$\mathbf{V}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2 .$$

Of particular importance: variance around a global (or local) minimum $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0 \Longrightarrow$ will be discussed later.

General idea to prove convergence

The general idea can be seen from the following decomposition:

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \\
&\quad = \mathbb{E}\|\mathbf{x}_k - \eta_k \nabla f_i(\mathbf{x}_k) - \mathbf{x}^*\|^2 \\
&\quad = \mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \underbrace{2\eta_k \mathbb{E}(\mathbf{x}_k - \mathbf{x}^*)^\top \nabla f_i(\mathbf{x}_k)}_{\text{Progress term}} + \underbrace{\eta_k^2 \mathbb{E}\|\nabla f_i(\mathbf{x}_k)\|^2}_{\text{Variance term}}.
\end{aligned}$$

Progress term: If f is strongly-convex, use $f(\mathbf{x}^*) - f(\mathbf{x}) \geq \nabla f(\mathbf{x}) \cdot (\mathbf{x}^* - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}\|^2$.

Variance: We need to **control the variance term** in order to ensure convergence. How?

1. Averaging iterates
2. Use appropriate decreasing step-size
3. Use an explicit mechanism to reduce the variance

Convergence for smooth function and bounded gradients

Theorem 9

Assume that the gradient of function f is L-Lipschitz and the stochastic gradients are bounded by a constant σ. Let $\eta_k = \frac{c}{\sqrt{K}}$ with $c = \sqrt{\frac{2(f(\mathbf{x}_0)-f^*)}{L\sigma^2}}$, then the iterates of SGD satisfy

$$\min_{s\in[0,K-1]} \mathbb{E}\left\|\nabla f(\mathbf{x}_s)\right\|^2 \leq \sqrt{\frac{2(f(\mathbf{x}_0)-f^*)L}{K}}\sigma \tag{2}$$

Section 5

Momentum and Adaptivity

Nesterov's Accelerated Method

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \beta(\mathbf{x}_k - \mathbf{x}_{k-1})$$

$$\mathbf{x}_{k+1} = \mathbf{y}_{k+1} - \eta \nabla f(\mathbf{y}_{k+1})$$

1. $\beta \geq 0$ is a momentum parameter.
2. $\beta = 0$ equals standard gradient descent.
3. Extrapolate iterates $(\mathbf{y}^k)$
4. Proceed from extrapolated point with an extra gradient step.

Theorem: Acceleration

Theorem 10 (Nesterov 2004)

Let f be L-smooth and μ-strongly convex, $\kappa = \frac{L}{\mu}$. Then with the choice $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ it holds that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq L \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa}} \right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

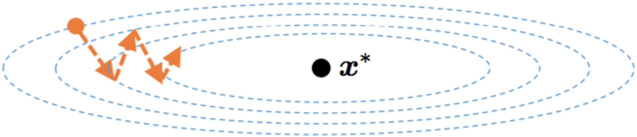
1. Improves dependency on the **condition number κ.**
2. Matches up to constants tightest known lower-cases bound.
3. Accelerated GD is essentially **optimal** for this class of functions.
4. w/o strong convexity, acceleration from $\mathbf{O}(1/k)$ to $\mathbf{O}(1/k^2)$.

Polyak's Heavy Ball Method

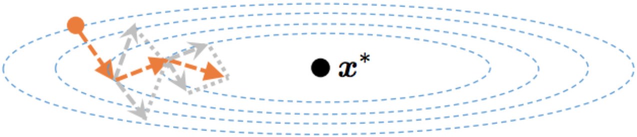
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k) + \beta(\mathbf{x}_k - \mathbf{x}_{k-1}), \quad \beta \in (0;1)$$

1. Often used in practice.
2. Easily derived from physical system dynamics.
3. Accelerated convergence not fully understood. (But: locally)
4. Certainly can accelerate the transient phase of gradient descent

Heavy Ball: Illustration



gradient descent



heavy-ball method

(Taken from `http://www.princeton.edu/~yc5/ele522_optimization/lectures/accelerated_gradient.pdf`)

Polyak's Heavy Ball Method

What if the gradient does not change (much)?

$$
\begin{aligned}
\mathbf{x}_1 - \mathbf{x}_0 &= -\eta \nabla f \\
\mathbf{x}_2 - \mathbf{x}_1 &= -\eta (1 + \beta) \nabla f \\
\mathbf{x}_3 - \mathbf{x}_2 &= -\eta (1 + \beta(1 + \beta)) \nabla f = -\eta (1 + \beta + \beta^2) \nabla f \\
&\dots
\end{aligned}
$$

So that in the limit

$$
\lim_{k \to \infty} (\mathbf{x}_k - \mathbf{x}_{k-1}) = -\eta \nabla f \sum_{i=0}^{\infty} \beta^i = -\eta \nabla f \left[\frac{1}{1 - \beta} \right]
$$

$\beta = 0.9$ results (potentially) in 10-fold acceleration.

AdaGrad: Motivation

Adapt learning rate per parameter or dimension.

Originally: motivated from sparse features (e.g. text)

Compositional models: adapted step size for different parameters (in different layers)

AdaGrad

Define the monotonically increasing sequence

$$\boldsymbol{\gamma}^k = \boldsymbol{\gamma}^{k-1} + \nabla f(\mathbf{x}_k) \odot \nabla f(\mathbf{x}_k)$$

where $\odot$ denotes pointwise multiplication.

γ_i^k corresponds to the sum of the squares of the i-th parameter's partial derivatives along the iterates $\mathbf{x}_0, \ldots, \mathbf{x}_k$.

AdaGrad

Use these estimates as diagonal pre-conditioner

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\, \mathbf{\Lambda}^k\, \nabla f(\mathbf{x}_k), \quad \text{where}$$

$$\mathbf{\Lambda}^k = \text{diag}(\lambda_i^k), \text{ with } \lambda_i^k = \frac{1}{\sqrt{\gamma_i^k} + \delta}, \quad \delta > 0$$

- sophisticated convergence analysis w/ regret bounds
- ... beyond our scope.

Adam: Momentum

Exponentially decaying mean over the gradient history

$$\mathbf{m}^k = \beta\, \mathbf{m}^{k-1} + (1-\beta)\, \nabla f(\mathbf{x}_k), \quad \beta \in [0;1], \quad \mathbf{m}^0 = \mathbf{0}$$

Momentum "with friction":

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\eta}{1-\beta^k}\, \mathbf{m}^k\,. \tag{3}$$

- denominator: bias correction (zero initialization)
- different dynamics than heavyball

Adam: Adaptivity

Exponentially decaying average of uncentered variances

$$\boldsymbol{\gamma}^k = \alpha \boldsymbol{\gamma}^{k-1} + (1-\alpha)\left[\nabla f(\mathbf{x}_k) \odot \nabla f(\mathbf{x}_k)\right]$$

Diagonal pre-conditioner with momentum term = **Adam**

$$\mathbf{x}_k + 1 = \mathbf{x}_k - \frac{\eta}{1-\beta^k}\, \boldsymbol{\Lambda}^k\, \mathbf{m}^k$$

$$\boldsymbol{\Lambda}^k = \text{diag}(\lambda_i^k), \quad \text{with} \quad \frac{1}{\lambda_i^k} = \sqrt{\frac{\gamma_i^k}{1-\alpha^k}} + \delta$$

- typical choices are $\beta = 0.9$ and $\alpha = 0.999$
- no invariance re:parameterization

AMSGrad

$$\hat{\gamma}^{k+1} = \max\left\{\hat{\gamma}^{k}, \gamma^{k}\right\}$$

- fixes a problem that may prevent Adam from converging (even on convex functions)