Deep Learning

Lecture 4: Non-convex Optimization for Learning (Part 1)

University of Agder,
Kristiansand, Norway

Prepared by: Hadi Ghauch⋆, Hossein S. Ghadikolaei†

⋆ Telecom ParisTech

† Royal Institute of Technology, KTH

https://sites.google.com/view/fundl/home

Outline

1. Motivation

2. Optimization of problems without structure
   - First-order methods
   - Successive Approximation

3. Optimization of problems with structure

Outline

Recap of Convex Optimization Algorithm

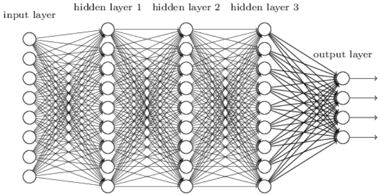
Strongly convex optimization problem: $\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i \in [N]} f_i(\boldsymbol{w}) + r(\boldsymbol{w})$

Existence of global optimal and efficient solution methods

Deterministic algorithms (Lec 2)
Gradient descent: move along steepest descent (complexity linear in N)
Newton's methods: include Hessian (curvature) information in the update. Not used in learning (complexity cubic in d)
Nesterov Acceleration: mix between interpolation and gradient descent. Need to know, L (complexity linear in N)
Momentum Methods (Heavy ball method): add momentum term to reduce the oscillations and improve convergence. Need to know, L, μ (complexity linear in N)
Stochastic Gradient Descent
-computational complexity independent of N
SGD family for smooth problems
noise reduction techniques and adaptive mini-batches, SVRG, and SAGA

Motivation

Resurgence of AI is due to **Deep Neural Networks (DNNs)**
and countless variants (covered later)



DNNs is a composition of **non-linear layers**, $\boldsymbol{W}_1, ..., \boldsymbol{W}_J$:

$$\boldsymbol{y}_i = \sigma_J(\boldsymbol{W}_J \cdots \sigma_2(\boldsymbol{W}_2 \sigma_1(\boldsymbol{W}_1 \boldsymbol{x}_i))),\ i \in [N]$$

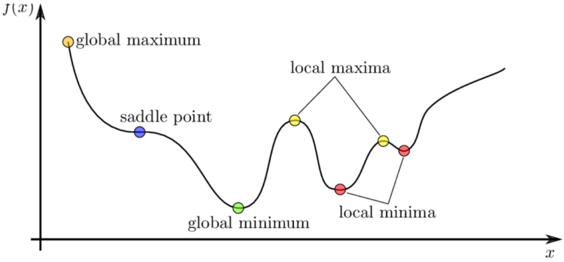
Let $\boldsymbol{w} \in \mathbb{R}^d$ be the total number of weights in DNN (from all layers)
$d \geq 10^6$ in current deep learning application

The resulting optimization DNN training

$$\min_{\boldsymbol{w} \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} f_i((\boldsymbol{x}_i, \boldsymbol{y}_i); \boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2 := f(\boldsymbol{w})$$

non-convex optimization problem

Definitions



$\boldsymbol{w}^\star$ **global minimum** iff $\nabla f(\boldsymbol{w}^\star) = 0,\ \nabla^2 f(\boldsymbol{w}^\star) \succeq 0, f(\boldsymbol{w}^\star) \leq f(\boldsymbol{w}),\ \forall\ \boldsymbol{w} \in \mathcal{D}$

$\boldsymbol{w}^\star$ **local minimum** iff $\nabla f(\boldsymbol{w}^\star) = 0,\ \nabla^2 f(\boldsymbol{w}^\star) \succeq 0, f(\boldsymbol{w}^\star) \leq f(\boldsymbol{w}),\ \forall\ \boldsymbol{w} \in \mathcal{B}$

$\boldsymbol{w}^\star$ **non-degenerate saddle points** iff $\nabla f(\boldsymbol{w}^\star) = 0,\ \nabla^2 f(\boldsymbol{w}^\star)$ has strictly positive and negative eigenvalues

$\boldsymbol{w}^\star$ **saddle point** iff $\nabla f(\boldsymbol{w}^\star) = 0,\ \nabla^2 f(\boldsymbol{w}^\star)$ has (positive and negative eigenvalues)

$\boldsymbol{w}^\star$ **stationary point** iff $\nabla f(\boldsymbol{w}^\star) = 0,\ \boldsymbol{w}^\star \in \mathcal{D}$

decrease in level of restrictiveness

Definitions

Problem: $\arg\min_{\boldsymbol{w}\in\mathcal{D}} f(\boldsymbol{w})$

Convex world: any point $\nabla f(\boldsymbol{w}_k) = 0, \boldsymbol{w}_k \in \mathcal{D}$ globally optimal point

Nononvex world: any point $\nabla f(\boldsymbol{w}_k) = 0, \boldsymbol{w}_k \in \mathcal{D}$ is stationary point:
global min ? local min ? saddle point ?
global solution hard for general non-convex prob
Goal: bypass stationary points (usually bad) and find local minimum

Second-order necessary (2oN) point: $\|\nabla f(\boldsymbol{w}_k)\|_2^2 = 0$ & $\nabla^2 f(\boldsymbol{w}_k) \succeq \mathbf{0}$
- point with zero gradient and positive curvature: local minimum

Approximate 2oN point: $\|\nabla f(\boldsymbol{w}_k)\|_2^2 \leq \epsilon_g$, $\nabla^2 f(\boldsymbol{w}_k) \succeq -\epsilon_H \boldsymbol{I}$ for small positive ϵ_g, ϵ_H
- approximation of local minimum (relaxation of conditions of 2oN pt)

Complexity measure:

gradient evaluations: # calls to incremental first-order oracle with input $(\boldsymbol{w}, i)$ and output $(f_i(\boldsymbol{w}), \nabla f_i(\boldsymbol{w}))$

Basic assumptions

f is bounded below: $f(\boldsymbol{w}) \geq f_{\inf}$ for all $\boldsymbol{w} \in \mathcal{D}$

Function, its gradient and Hessian are Lipschitz continuous

$$\|f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1)\|_2 \leq L\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$$
$$\|\nabla f(\boldsymbol{w}_2) - \nabla f(\boldsymbol{w}_1)\|_2 \leq L_g\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$$
$$\|\nabla^2 f(\boldsymbol{w}_2) - \nabla^2 f(\boldsymbol{w}_1)\|_F \leq L_H\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$$

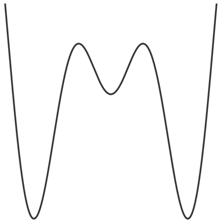
L, L_g and L_H are Lipschitz const of the function, the gradient and Hessian

Quadratic upper-bounds from Taylor's theorem ($\forall \boldsymbol{v}, \boldsymbol{w} \in \mathcal{D}$)
- approximate func at $f(\boldsymbol{w}+\boldsymbol{v})$ with Talor, using at grad and Hessian at $\boldsymbol{w}$:

$$f(\boldsymbol{w}+\boldsymbol{v}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^T \boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T \nabla^2 f(\boldsymbol{w}) \boldsymbol{v}$$

Non-convex Optimization

$$\min_{\boldsymbol{w}\in\mathcal{D}} f(\boldsymbol{w}) \qquad (1)$$



$f(\boldsymbol{w}) : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}$ is non-convex
$\mathcal{D}$: domain of f (convex or non-convex)

Local optimality may not necessarily imply global optimality

Proper initialization is very important in nonconvex optimization

First-order criteria $(\nabla f(\boldsymbol{w}) = 0)$
 necessary and sufficient conditions for convex
 only necessary condition for nonconvex

Strategies for Non-convex Optimization

No generic method for all non-convex problems ($\neq$ convex case)
solution approach depends on **structure of the optimization**

1. **No structure on $f(w)$:**
   first-order methods may be used. convergence?
   successive approximation: successively approx (1) with linear/convex bound

2. $f(\boldsymbol{w})$ **is coordinate separable** into convex subproblems (k-means, Lyods algo)
   if $f(\boldsymbol{w}) = f(w_1, ..., w_d)$, where $(w_1, ..., w_d)$ **coordinates**
   solve (1) as series **subproblems**: alternatively min each **subproblems** while fixing all other ones. Strong convexity of subproblems to prove convergence

3. $f(\boldsymbol{w})$ **is block-coordinate separable** into convex subproblems:
   if $f(\boldsymbol{w}) = f(\boldsymbol{w}_1, ..., \boldsymbol{w}_d)$, where $(\boldsymbol{w}_1, ..., \boldsymbol{w}_d)$ **block of coordinates**
   solve (1) as series **subproblems**: alternatively min each **subproblems** while fixing all other ones. Strong convexity of subproblems to prove convergence

4. $f(\boldsymbol{w})$ **is block-coordinate separable** into non-convex subproblems:
   solve (1) as series **subproblems**: alternatively min each **subproblems** while fixing all other ones. Strong convexity of subproblems not needed to prove convergence

Outline

Can we use a simple GD?

One-point convexity:

Convexity implies $f(\boldsymbol{w}_k) - f(\boldsymbol{v}) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{v})$ for all $\boldsymbol{v}$

Then it must hold for a local min: $\boldsymbol{v} = \boldsymbol{w}^\star$?

What if $w^\star$ is global min? GD converges to global min

Run a GD step: $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$.
Use def of convexity show $\Rightarrow \Delta_k := f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$
f **convex only in neighborhood of** $w^\star$**:** $\Delta_k \leq \mathcal{O}(1/k)$

Linear convergence to local min.
computational complexity: $\mathcal{O}(N)$ gradient evaluations per iteration

assuming one point convexity: func is locally convex around some local min

This condition holds for many functions

Challenge for GD in non-convex setting?

poor scalability with N

convergence to a stationary point, not necessarily a local minimum

results conditioned on showing f satisfies one point convexity condition

Can we use a simple GD?

One-point convexity:

Convexity implies $f(\boldsymbol{w}_k) - f(\boldsymbol{v}) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{v})$ for all $\boldsymbol{v}$

Then it must hold for a local min: $\boldsymbol{v} = \boldsymbol{w}^\star$?

What if $w^\star$ is global min? GD converges to global min

Run a GD step: $w_{k+1} = w_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$.
Use def of convexity show $\Rightarrow \Delta_k := f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$
f **convex only in neighborhood of** $w^\star$**:** $\Delta_k \leq \mathcal{O}(1/k)$

Linear convergence to local min.
computational complexity: $\mathcal{O}(N)$ gradient evaluations per iteration

assuming one point convexity: func is locally convex around some local min

This condition holds for many functions

Challenge for GD in non-convex setting?

poor scalability with N

convergence to a stationary point, not necessarily a local minimum

results conditioned on showing f satisfies one point convexity condition

Can we use a simple GD?

One-point convexity:

Convexity implies $f(\boldsymbol{w}_k) - f(\boldsymbol{v}) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{v})$ for all $\boldsymbol{v}$

Then it must hold for a local min: $\boldsymbol{v} = \boldsymbol{w}^\star$?

What if $w^\star$ is global min? GD converges to global min

Run a GD step: $w_{k+1} = w_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$.
Use def of convexity show $\Rightarrow \Delta_k := f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$
f **convex only in neighborhood of** $w^\star$**:** $\Delta_k \leq \mathcal{O}(1/k)$

Linear convergence to local min.
computational complexity: $\mathcal{O}(N)$ gradient evaluations per iteration

assuming one point convexity: func is locally convex around some local min

This condition holds for many functions

Challenge for GD in non-convex setting?

- poor scalability with N
- convergence to a stationary point, not necessarily a local minimum

results conditioned on showing f satisfies one point convexity condition

Can we use a simple GD?

One-point convexity:

Convexity implies $f(\boldsymbol{w}_k) - f(\boldsymbol{v}) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{v})$ for all $\boldsymbol{v}$

Then it must hold for a local min: $\boldsymbol{v} = \boldsymbol{w}^\star$?

What if $w^\star$ is global min? GD converges to global min

Run a GD step: $w_{k+1} = \boldsymbol{w}_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$.
Use def of convexity show $\Rightarrow \Delta_k := f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$
f **convex only in neighborhood of** $w^\star$**:** $\Delta_k \leq \mathcal{O}(1/k)$

Linear convergence to local min.
computational complexity: $\mathcal{O}(N)$ gradient evaluations per iteration

assuming one point convexity: func is locally convex around some local min

This condition holds for many functions

Challenge for GD in non-convex setting?

poor scalability with N

convergence to a stationary point, not necessarily a local minimum

results conditioned on showing f satisfies one point convexity condition

One-point convexity [Allen-Zhu, ICML, 2017]

Convex f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq f(\boldsymbol{w}) - f(\boldsymbol{v})$$

μ-strongly convex

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \mu \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star) \right)$$

μ-strongly convex and L_g-smooth

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \frac{1}{2L_g} \|\nabla f(\boldsymbol{w})\|_2^2$$

If f **nonconvex** but satisfies

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star) \right)$$

If f **nonconvex** but satisfies

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

two-layer NN [Li-Yuan, 2017]

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star) \right)$$

(known as Polyak-Lojasiewicz condition)

finite sum minimization [Reddi-Sra-Poczos-Smola, 2016]

If f **nonconvex** but satisfies

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \gamma \|\nabla f(\boldsymbol{w})\|_2^2$$

dictionary learning [Arora-Ge-Ma-Moitra, 2015]

One-point convexity [Allen-Zhu, ICML, 2017]

Convex f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq f(\boldsymbol{w}) - f(\boldsymbol{v})$$

μ**-strongly convex**

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \mu \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

μ**-strongly convex and** L_g**-smooth**

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \frac{1}{2L_g} \|\nabla f(\boldsymbol{w})\|_2^2$$

If f **nonconvex** but satisfies

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

If f **nonconvex** but satisfies

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

two-layer NN [Li-Yuan, 2017]

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

(known as Polyak-Lojasiewicz condition)

finite sum minimization [Reddi-Sra-Poczos-Smola, 2016]

If f **nonconvex** but satisfies

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \gamma \|\nabla f(\boldsymbol{w})\|_2^2$$

dictionary learning [Arora-Ge-Ma-Moitra, 2015]

One-point convexity, how to solve?

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq f(\boldsymbol{w}) - f(\boldsymbol{v})$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \mu \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \frac{1}{2L_g} \|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \gamma \|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

One-point convexity, how to solve?

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq f(\boldsymbol{w}) - f(\boldsymbol{v})$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \mu \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \frac{1}{2L_g}\|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \gamma\|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

One-point convexity, how to solve?

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq f(\boldsymbol{w}) - f(\boldsymbol{v})$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \mu \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \frac{1}{2L_g} \|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \gamma \|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

One-point convexity, how to solve?

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{v} \rangle \geq f(\boldsymbol{w}) - f(\boldsymbol{v})$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \mu \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \frac{1}{2L_g} \|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

GD/SGD converges in $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
GD converges in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ for smooth f

$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \beta \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

GD/SGD converges in $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$

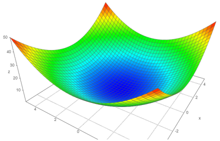
$$\|\nabla f(\boldsymbol{w})\|_2^2 \geq \beta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\right)$$

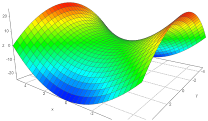
GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$ for smooth f

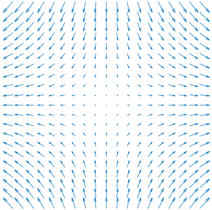
$$\langle \nabla f(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \rangle \geq \frac{\beta}{2} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + \gamma \|\nabla f(\boldsymbol{w})\|_2^2$$

GD converges in $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$

How to escape non-degenerate saddle points







Find an approximate 2oN point: $\|f(\boldsymbol{w}_k)\|_2^2 \leq \epsilon_g, \nabla^2 f(\boldsymbol{w}_k) \succeq -\epsilon_H \boldsymbol{I}$
- approximate local min (zero gradient with positive curvature)

Gradient-based
1) Perturbed GD: $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k) + \text{noise}$
$\log(d)$ complexity on the parameter size [Jin-Ge-Netrapalli-Kakade-Jordan, 2017]

2) SGD

Newton's method: too expensive

Hessian-based
Faster reaction to saddle points using curvature information, expensive iterations could be very efficient [Allen-Zhu, 2018]

A generic Hessian-based algorithm

1. If $\|\nabla f(\boldsymbol{w}_k)\|_2 > \epsilon_g$, run your favorite algorithm (say GD with step-size $1/L_g$ to get closer to a stationary point
2. else, if $\nabla^2 f(\boldsymbol{w}_k) \not\succeq \epsilon_h \boldsymbol{I}$ there is at least one negative eigenval
   find the eigenvector, $\boldsymbol{v}$, of the smallest eigenvalue ($\lambda_{\min}^k$) of $\nabla^2 f(\boldsymbol{w}_k)$, namely
   $$\|\boldsymbol{v}_k\|_2^2 = 1, \quad \boldsymbol{v}^T \nabla^2 f(\boldsymbol{w}_k)\boldsymbol{v} = \lambda_{\min}^k, \quad \nabla f(\boldsymbol{w}_k)^T \boldsymbol{v} \leq 0$$
3. Move toward that direction. why? $\boldsymbol{v}$ is a descent direction, since $\nabla f(\boldsymbol{w}_k)^T \boldsymbol{v} \leq 0$
   $$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \frac{2|\lambda_{\min}^k|}{L_H}\boldsymbol{v}$$
4. Otherwise, terminate.

$\Rightarrow$ The number of iterations is at most $\max\left(\frac{2L_g}{\epsilon_g^2}, \frac{3L_H^2}{2\epsilon_H^3}\right)(f(\boldsymbol{w}_0) - f_{\inf})$

Notice from Taylor expansion that $f(\boldsymbol{w}_k + \boldsymbol{v}) \approx f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w})^T \boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{v}$. Step 2 finds a direction (v) that gives the highest reduction to f. Do we need to really find the minimum eigenvalue? or a strong negative one is enough?

A generic Hessian-based algorithm

1. If $\|\nabla f(\boldsymbol{w}_k)\|_2 > \epsilon_g$, run your favorite algorithm (say GD with step-size $1/L_g$ to get closer to a stationary point

2. else, if $\nabla^2 f(\boldsymbol{w}_k) \not\succeq \epsilon_h \boldsymbol{I}$ there is at least one negative eigenval
   find the eigenvector, $\boldsymbol{v}$, of the smallest eigenvalue ($\lambda_{\min}^k$) of $\nabla^2 f(\boldsymbol{w}_k)$, namely

$$\|\boldsymbol{v}_k\|_2^2 = 1, \quad \boldsymbol{v}^T \nabla^2 f(\boldsymbol{w}_k)\boldsymbol{v} = \lambda_{\min}^k, \quad \nabla f(\boldsymbol{w}_k)^T \boldsymbol{v} \leq 0$$

3. Move toward that direction. why? $\boldsymbol{v}$ is a descent direction, since $\nabla f(\boldsymbol{w}_k)^T \boldsymbol{v} \leq 0$

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \frac{2|\lambda_{\min}^k|}{L_H}\boldsymbol{v}$$

4. Otherwise, terminate.

$\Rightarrow$ The number of iterations is at most $\max\left(\frac{2L_g}{\epsilon_g^2}, \frac{3L_H^2}{2\epsilon_H^3}\right)(f(\boldsymbol{w}_0) - f_{\inf})$

Notice from Taylor expansion that $f(\boldsymbol{w}_k + \boldsymbol{v}) \approx f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w})^T \boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{v}$. Step 2 finds a direction ($\boldsymbol{v}$) that gives the highest reduction to f. Do we need to really find the minimum eigenvalue? or a strong negative one is enough?

Outline

1. Motivation

2. Optimization of problems without structure
   - First-order methods
   - Successive Approximation

3. Optimization of problems with structure

Successive Approximation Methods

$$\arg\min_{\boldsymbol{w}\in\mathcal{D}} f(\boldsymbol{w})$$

Idea: minimize an **approximate surrogate** fnc.
Update approximation iteratively to make it locally tight:

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}\in\mathcal{D}} \widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k)$$

- $\widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k)$ is the approximation of f at $\boldsymbol{w}_k$

Two intuitive choices for $\widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k)$:

- **Gradient**: Successive Linear Approximation
- **Hessian**: Successive Convex Approximation

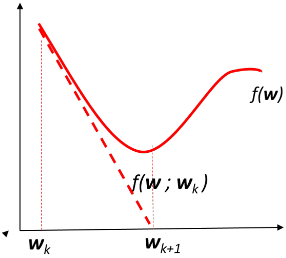
Successive Linear Approximation

Successive Linear Approximation (SLA)

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w} \in \mathcal{D}} \widetilde{f}(\boldsymbol{w}; \boldsymbol{w}_k) \qquad (2)$$

$\widetilde{f}(\boldsymbol{w}; \boldsymbol{w}_k)$ is first order approx of f at $\boldsymbol{w}_k$:

$$\widetilde{f}(\boldsymbol{w}; \boldsymbol{w}_k) := f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^T (\boldsymbol{w} - \boldsymbol{w}_k) + (\gamma/2) \, \|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2$$



Theorem 1: suppose that f is differentiable and Lipschitz continuous. Thus, the updates in (2) converge to a **stationary point** of f, as $k \to \infty$

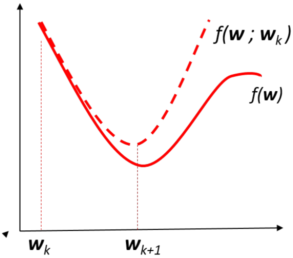
Successive Convex Approximation

Successive Convex Approximation (SCA)

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}\in\mathcal{D}} \widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k) \qquad (3)$$

$\widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k)$ is convex approx of f at $\boldsymbol{w}_k$:

$$\begin{aligned}\widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k) := f(\boldsymbol{w}_k) &+ \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}-\boldsymbol{w}_k) \\ &+ (\boldsymbol{w}-\boldsymbol{w}_k)^T\nabla^2 f(\boldsymbol{w}_k)(\boldsymbol{w}-\boldsymbol{w}_k)\end{aligned}$$



Theorem 2: suppose that f is differentiable and Lipschitz continuous. Thus, the updates in (3) converge to a **stationary point** of f, as $k \to \infty$

Successive Approximation Methods

Take-home messages:

- In general, **no guarantees** on global convergence
  only convergence to stationary point
- Quality of solution depends on initial point
- SCA converges quadratically, SLA converges linearly
- Computational complexity too high for SCA: $\mathcal{O}(d^3)$
  from inverting Hessian matrix

Outline

1. Motivation

2. Optimization of problems without structure
   First-order methods
   Successive Approximation

3. Optimization of problems with structure

Formulation into equivalent problem

Some nonconvex problems are **equivalently reformulated** into convex form.
-see equivalence between optimization problems (Chapter 3, [Boyd, 2004])

Rayleigh-Ritz quotient:
basis of **linear discriminant analysis** in dimensionality reduction
given **two classes of points** generated by two distributions with cov matrices, where $\boldsymbol{A} \succeq \mathbf{0}, \boldsymbol{B} \succ \mathbf{0}$
find the hyperplane, $\boldsymbol{w}$, maximizing expected dist among two classes after projection on $\boldsymbol{w}$

$$\max_{\|\boldsymbol{w}\|_2=1} \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{B} \boldsymbol{w}}$$

the optimization problem is equivalent to the min eigenvector. How ?

Minimum eigenvector of a symmetric matrix $\boldsymbol{A}$: $\min_{\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}$
Nonconvex in the Euclidean space. why ?
Globally optimal solution by equivalently formulating the problem on the Riemann manifold

In general no systematic approach (rather problem specific)

Conclusions

Solution method depends on structure of problem
Non-convex optimization w/out structure:

- first-order methods: with tweaks to escape saddle pts
- Successive linear/convex approximation

Non-convex optimization with structure: (Next)

Some references

- Stephen J. Wright, “Coordinate descent algorithm” , Math. Program., 2015.
- S. Reddi, S. Sra, B. Poczos, and A. Smola, “Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization,”, NIPS, 2016.
- S. ShalevShwartz and T. Zhang, “Stochastic dual coordinate ascent methods for regularized loss minimization,” JMLR, 2012.
- Y. Li and Y. Yuan, “Convergence analysis of two-layer neural networks with ReLU activation,”, NIPS, 2017.
- M. Razaviyayn, M. Hong, and Z.-Q. Luo, “A unified convergence analysis of block successive minimization methods for nonsmooth optimization”, SIAM J. Optim., 2013.
- M. Hong, M. Razaviyayn, Z. Q. Luo and J. S. Pang, “A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing,” IEEE SPM, 2016.
- Z. Allen-Zhu, “Recent advances in stochastic convex and non-convex optimization,” ICML Tutorial, 2017.
- S. Arora, R. Ge, T. Ma, and A Moitra, “Simple, efficient, and neural algorithms for sparse coding,” JMLR, 2015.
- C. Jin, R. Ge, P. Netrapalli, S.M. Kakade, and M.I. Jordan, “How to escape saddle points efficiently,” ICML, 2017.
- Z. Allen-Zhu, “Natasha 2: Faster non-convex optimization than SGD,” NIPS, 2018.
- Y. Zhang, P. Liang, and M. Charikar, “A hitting time analysis of stochastic gradient Langevin dynamics,” COLT, 217.

Deep Learning

Lecture 4: Non-convex Optimization for Learning (Part 1)

University of Agder,
Kristiansand, Norway

Prepared by: Hadi Ghauch*, Hossein S. Ghadikolaei†

* Telecom ParisTech

† Royal Institute of Technology, KTH

https://sites.google.com/view/fundl/home