ROYAL INSTITUTE
OF TECHNOLOGY

EP3260: Machine Learning Over Networks

Lecture 4: Centralized Nonconvex ML

Hossein S. Ghadikolaei, Hadi Ghauch, and Carlo Fischione

Division of Network and Systems Engineering
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology, Stockholm, Sweden

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

February 2019

Learning outcomes

- Recap of stochastic iterative algorithms for convex optimization
- Hardness of nonconvex problems
- Finding stationary points for (non-)smooth problems
- Escaping saddle points
- Finding an approximate local minima
- Convergence results for large-scale nonconvex optimization

Outline

1. Nonconvex optimization
2. Finding stationary points for structured nonconvex problems
3. Finding stationary points for generic nonconvex problems
4. Finding a local minima
5. Supplements

Recap of convex solvers

Our main optimization problem: $\underset{\boldsymbol{w}}{\text{minimize}}\ \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) + r(\boldsymbol{w})$

Existence of global optimality and efficient solvers

Deterministic solution algorithms

-gradient Oracle's load is linear with N

GD family for smooth problems

subgradient and proximal methods for non-smooth (or composite) functions

Stochastic solution algorithms

-gradient Oracle's load is independent of N

SGD family for smooth problems

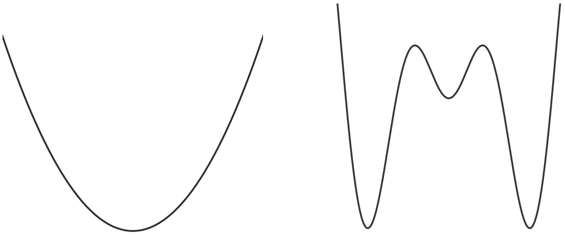
noise reduction techniques and adaptive mini-batches, SVRG, and SAGA

proximal methods for non-smooth (or composite) functions

Outline

1. Nonconvex optimization
2. Finding stationary points for structured nonconvex problems
3. Finding stationary points for generic nonconvex problems
4. Finding a local minima
5. Supplements

Nonconvexity



Nonconvex optimization can encode most problems

Local optimality may not necessarily imply global optimality

Proper initialization is very important in nonconvex optimization

First-order criteria ($\|\nabla f(\boldsymbol{w})\|_2 \to 0$)

- necessary and sufficient conditions for convex
- only necessary condition for nonconvex

Nonconvex optimization

Subset-sum problem: find a subset of $\{a_1, a_2, \ldots, a_N\}$ that sums to b. Lets formulate it as a nonconvex optimization problem:

$$\underset{\boldsymbol{w} \in \{0,1\}^N}{\text{minimize}} \; \left(\boldsymbol{a}^T \boldsymbol{w} - b\right)^2 + \boldsymbol{w}^T (\mathbf{1} - \boldsymbol{w})$$

Can we achieve the global minima (e.g., 0)?

It is NP-complete

OK! give up the global optimal point. Can we run GD to find a local minima?

Curse of dimensionality: finding a local optima becomes exponentially harder

$\Rightarrow$ Results in nonconvex optimization are not as strong as the convex case

Nonconvex optimization

Subset-sum problem: find a subset of $\{a_1, a_2, \ldots, a_N\}$ that sums to b. Lets formulate it as a nonconvex optimization problem:

$$\underset{\boldsymbol{w} \in \{0,1\}^N}{\text{minimize}} \ \left(\boldsymbol{a}^T \boldsymbol{w} - b\right)^2 + \boldsymbol{w}^T (\boldsymbol{1} - \boldsymbol{w})$$

Can we achieve the global minima (e.g., 0)?

It is NP-complete

OK! give up the global optimal point. Can we run GD to find a local minima?

Curse of dimensionality: finding a local optima becomes exponentially harder

$\Rightarrow$ Results in nonconvex optimization are not as strong as the convex case

Nonconvex optimization

Subset-sum problem: find a subset of $\{a_1, a_2, \ldots, a_N\}$ that sums to b.
Lets formulate it as a nonconvex optimization problem:

$$\underset{\boldsymbol{w} \in \{0,1\}^N}{\text{minimize}} \ \left(\boldsymbol{a}^T \boldsymbol{w} - b\right)^2 + \boldsymbol{w}^T (\mathbf{1} - \boldsymbol{w})$$

Can we achieve the global minima (e.g., 0)?

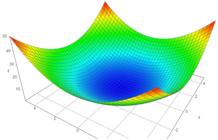
It is NP-complete

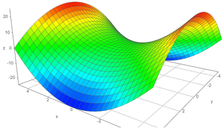
OK! give up the global optimal point. Can we run GD to find a local minima?

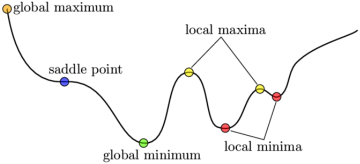
Curse of dimensionality: finding a local optima becomes exponentially harder

$\Rightarrow$ Results in nonconvex optimization are not as strong as the convex case

Some definitions







Stationary/critical points: $\{w \mid \nabla f(\boldsymbol{w}) = 0\}$

Local minima: a critical point where $\nabla^2 f(\boldsymbol{w}) > 0$

Local maxima: a critical point where $\nabla^2 f(\boldsymbol{w}) < 0$

Non-degenerate saddle points: a critical point where $\nabla^2 f(\boldsymbol{w})$ has strictly positive and negative eigenvalues

visualize gradient flow around a non-degenerate saddle point

Other types of saddle points (e.g., monkey saddle), usually harder to identify and treat!

More definitions

Problem: $\underset{\boldsymbol{w}\in\mathcal{W}}{\text{minimize}}\ f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w})$

Convex land: optimality gap $\mathbb{E}\left[\|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \to 0$

expectation w.r.t. potential randomness of the algorithm

Nononvex land: stationarity gap $\mathbb{E}\left[\|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \to 0$

Second-order necessary (2oN) point: $\|\nabla f(\boldsymbol{w}_k)\|_2^2 \to 0$ & $\nabla^2 f(\boldsymbol{w}_k) \geq 0$

Approximate 2oN point: $\|\nabla f(\boldsymbol{w}_k)\|_2^2 \leq \epsilon_g$, $\nabla^2 f(\boldsymbol{w}_k) \geq -\epsilon_H \boldsymbol{I}$ for small positive ϵ_g, ϵ_H

Complexity measure:

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

#Hessian-vector products

#alternations among subproblems

Basic assumptions

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

Gradient and Hessian are Lipschitz continuous

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

Quadratic and cubic upper-bounds from Taylor's theorem ($\forall \boldsymbol{v}, \boldsymbol{w} \in \mathcal{W}$)

$$f(\boldsymbol{w} + \boldsymbol{v}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^T \boldsymbol{v} + \frac{L_g}{2}\|\boldsymbol{v}\|_2^2$$
$$f(\boldsymbol{w} + \boldsymbol{v}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^T \boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} + \frac{L_H}{6}\|\boldsymbol{v}\|_2^3$$

Outline

1. Nonconvex optimization
2. Finding stationary points for structured nonconvex problems
3. Finding stationary points for generic nonconvex problems
4. Finding a local minima
5. Supplements

Reformulation to something similar

- Minimum eigenvalue of a symmetric matrix $\boldsymbol{A}$: $\underset{\|\boldsymbol{w}\|_2=1}{\text{minimize}}\ \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}$

  Nonconvex on the Euclidean space, but easily solvable on the Riemannian manifolds [Smith, 1994]

  Run your favorite convex solver on the reformulated convex problem

- Quasi-convex problems, invex problems, etc.

- Usually ad-hoc, no generic approach, problem specific

[Smith, 1994] S.T. Smith, "Optimization techniques on Riemannian manifolds," Fields Institute Communications, 1994.

Can we use a simple GD?

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

What if we have that inequality only for one point: $\boldsymbol{v} = \boldsymbol{w}^\star$?

This condition holds for most functions in the vicinity of any *local minima* $\boldsymbol{w}^\star$

What if $w^\star$ is global minima?

Run $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$ and let $f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$

$\mathcal{O}(N)$ gradient oracle calls per iteration

$\mathcal{O}(L_g/\epsilon_g)$ iterations for L_g-smooth functions, so $\mathcal{O}(NL_g/\epsilon_g)$ calls (proof?)

So why not keep using GD even for nonconvex?

poor scalability with N

convergence to a stationary point, not necessarily a local minima

Can we use a simple GD?

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

What if we have that inequality only for one point: $\boldsymbol{v} = \boldsymbol{w}^\star$?

This condition holds for most functions in the vicinity of any *local minima* $\boldsymbol{w}^\star$

What if $w^\star$ is global minima?

Run $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$ and let $f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$

$\mathcal{O}(N)$ gradient oracle calls per iteration

$\mathcal{O}(L_g/\epsilon_g)$ iterations for L_g-smooth functions, so $\mathcal{O}(NL_g/\epsilon_g)$ calls (proof?)

So why not keep using GD even for nonconvex?

poor scalability with N

convergence to a stationary point, not necessarily a local minima

Can we use a simple GD?

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

What if we have that inequality only for one point: $\boldsymbol{v} = \boldsymbol{w}^\star$?

This condition holds for most functions in the vicinity of any *local minima* $\boldsymbol{w}^\star$

What if $w^\star$ is global minima?

Run $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\nabla f(\boldsymbol{w}_k)}{L_g}$ **and let** $f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}_k - \boldsymbol{w}^\star)$

$\mathcal{O}(N)$ gradient oracle calls per iteration

$\mathcal{O}(L_g/\epsilon_g)$ iterations for L_g-smooth functions, so $\mathcal{O}(NL_g/\epsilon_g)$ calls (proof?)

So why not keep using GD even for nonconvex?

poor scalability with N

convergence to a stationary point, not necessarily a local minima

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**

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

What if

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

What if

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

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]

What if

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

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**

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

What if

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

What if

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

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]

What if

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

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

Outline

1. Nonconvex optimization
2. Finding stationary points for structured nonconvex problems
3. Finding stationary points for generic nonconvex problems
4. Finding a local minima
5. Supplements

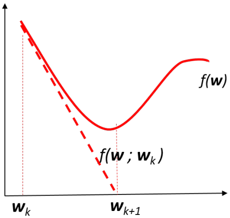
1. Successive approximation

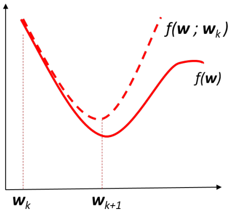
Minimize an (approximate) surrogate fnc. $\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k)$, where $\widetilde{f}(\boldsymbol{w};\boldsymbol{w}_k)$ is the approximation of f at $\boldsymbol{w}_k$

- successive linear approximation: $\widetilde{f} = f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}-\boldsymbol{w}_k)$
- successive quadratic approximation:

$$\widetilde{f} = f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^T(\boldsymbol{w}-\boldsymbol{w}_k) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_k)^T\nabla^2 f(\boldsymbol{w}_k)(\boldsymbol{w}-\boldsymbol{w}_k)$$

-GD acts as successive quadratic upper-bound minimization when we replace $\nabla^2 f(\boldsymbol{w}_k) \leq L\boldsymbol{I}$ (see Slide 2-25)



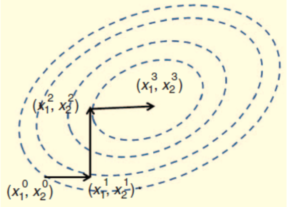


2. Coordinate descent (alternating methods)

Minimize only a block of coordinates while fixing others

$$\begin{aligned}\boldsymbol{w}_{(i)}^{k+1} &:= \arg\min_{\boldsymbol{w}_{(i)}} f(\boldsymbol{w}_{(1)}^k, \ldots, \boldsymbol{w}_{(i-1)}^k, \boldsymbol{w}_{(i)}, \boldsymbol{w}_{(i+1)}^k, \cdots, \boldsymbol{w}_{(d)}^k) \\ &= \arg\min_{\boldsymbol{w}_{(i)}} f(\boldsymbol{w}_{(i)}; \boldsymbol{w}^k)\end{aligned}$$

where $\boldsymbol{w}_{(i)}$ is i-th block of coordinates



[Hong-Razaviyayn-Luo-Pang, 2016]

Cyclic update rules, random coordinate selection, etc.

Convergence rate of $\mathcal{O}(dL_g/\epsilon_g)$ when each coordinate is L_g-smooth, namely $[\nabla^2 f(\boldsymbol{w})]_{ii} \in [0, L_g]$, convergence rate of $\mathcal{O}(dL_g/\mu \log \epsilon_g^{-1})$ for μ-strongly convex, [ShalevShwartz-Zhang, 2012]

What if the function is nonconvex/non-smooth in some coordinates?

3. Block successive upper-bound minimization

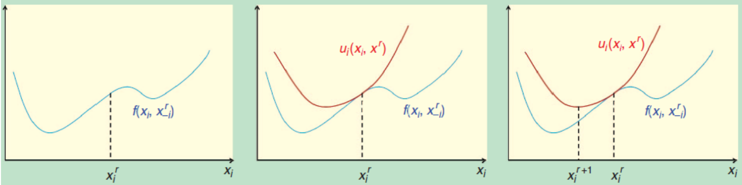
Combining block coordinate descent and successive convex approximation to handle non-convexity in $f(\boldsymbol{w}_{(i)}; \boldsymbol{w}^k)$

$$\boldsymbol{w}_{(i)}^{k+1} = \arg\min_{\boldsymbol{w}_{(i)}} \widetilde{f}(\boldsymbol{w}_{(i)}; \boldsymbol{w}^k)$$

where $\widetilde{f}$ is a strongly convex smooth upperbound of f at $\boldsymbol{w}^k$.

Example: proximal upper bound $\widetilde{f}(\boldsymbol{w}_{(i)}; \boldsymbol{w}^k) := f(\boldsymbol{w}_{(i)}; \boldsymbol{w}^k) + \frac{\gamma}{2}\|\boldsymbol{w}_{(i)} - \boldsymbol{w}_{(i)}^k\|_2^2$

Straightforward extension to composite functions ($f = g + h$ for smooth g and non-smooth h through proximal mapping)



[Hong-Razaviyayn-Luo-Pang, 2016]

Coordinate descent vs SGD

Coordinate Descent

Pick coordinates ζ_k and update $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k$ for coordinates $i \neq \zeta_k$ and

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla_i f(\boldsymbol{w}_k)$$

for coordinates $i = \zeta_k$

Compute all gradients for coordinates ζ_k

Guaranteed improvement in every iteration, consequently (usually) easier design and analysis

Stochastic Gradient Descent

Pick coordinate(s) ζ_k and update

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f_i(\boldsymbol{w}_k)$$

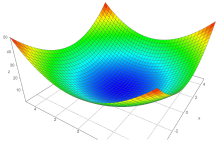
Compute one gradient f_{ζ_k} for all coordinates

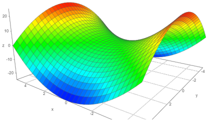
Not robust in general, need variance reduction techniques (like SVRG) to stabilize SGD

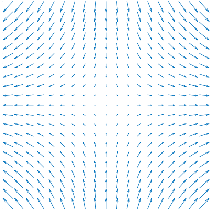
Outline

1. Nonconvex optimization
2. Finding stationary points for structured nonconvex problems
3. Finding stationary points for generic nonconvex problems
4. Finding a local minima
5. Supplements

How to escape non-degenerate saddle points







How to find an approximate 2oN point $(\|f(\boldsymbol{w}_k)\|_2^2 \leq \epsilon_g, \nabla^2 f(\boldsymbol{w}_k) \geq -\epsilon_H \boldsymbol{I})$?

Hessian-free approaches

1) Perturbed GD: $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k) + \text{noise}$

$\log(d)$ dependency on the parameter size [Jin-Ge-Netrapalli-Kakade-Jordan, 2017]

2) SGD

Hessian-based approaches

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. Otherwise, if $\nabla^2 f(\boldsymbol{w}_k) \not\succeq \epsilon_h \boldsymbol{I}$, find the most eigenvector of the most negative 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?), e.g., by

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

4. Otherwise, terminate.

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

Proof: see the board

[1] 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?

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. Otherwise, if $\nabla^2 f(\boldsymbol{w}_k) \not\succeq \epsilon_h \boldsymbol{I}$, find the most eigenvector of the most negative 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?), e.g., by

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

4. Otherwise, terminate.

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

Proof: see the board

[1] 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?

Foods for thought

1. Consider iterations of the acceleration methods developed for an L_g-smooth μ-strongly convex problem. How to use them in a smooth nonconvex setting?

   Estimate the minimum eigenvalue of $\nabla^2 f(\boldsymbol{w}_k)$, add a proper convex quadratic term to the objective.

   Find L_g and μ for this new objective function and run Nestrov's iterations. Do we converge? at which rate? What about a non-smooth nonconvex objective?

2. Lanczos algorithm efficiently finds the smallest eigenvalue of a symmetric matrix $\boldsymbol{A}$. Check the definition of Krylov subspace, generated by $\boldsymbol{A}$. Find a vector in the Krylov subspace that minimizes $\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z}$. Show that you can control the accuracy of the approximated eigenvalue by changing the degree of the Krylov subspace.

3. In our examples of successive convex approximation, we have used a *low-order* Taylor approximation around $\boldsymbol{w}_k$. Can we trust this approximation anyway?

   What about finding minimizer over a region to which we can trust? Define a trust region as a ball centered at $\boldsymbol{w}_k$ with radius Δ_k.

   Now, modify the successive quadratic approximation of slide 4-16 to find $\boldsymbol{v}$ in our trusted region. Adjust Δ_k to ensure a sufficient decrease in f at each iteration

   Congratulation! you have discovered the famous Trust Region method!

   Extend to cubic regularization by replacing the 3rd term of the Taylor expansion by $M_k \|\boldsymbol{v}\|^3$ for some positive M_k (Nestrov and Polyak, 2006)? Observe that $M_k \geq L_h/6$ leads to a successive upper-bound minimization.

4. Check `http://www.offconvex.org` and `https://www.facebook.com/nonconvex`

Which algorithm to choose?

CA3: Training a deep neural network

Consider optimization problem

$$\underset{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{w}_3}{\text{minimize}} \frac{1}{N} \sum_{i \in [N]} \|\boldsymbol{w}_3 \boldsymbol{s}(\boldsymbol{W}_2 \boldsymbol{s}(\boldsymbol{W}_1 \boldsymbol{x}_i) - \boldsymbol{y}_i\|_2^2,$$

where $\boldsymbol{s}(\boldsymbol{x}) = 1/(1 + \exp(-\boldsymbol{x}))$. You may add your choice of regularizer.

Consider both "Communities and crime" and "Individual household electric power consumption" regression datasets.

1) Try to solve this optimization task with proper choices of size of decision variables (matrix $\boldsymbol{W}_1$, matrix $\boldsymbol{W}_2$, and vector $\boldsymbol{w}_3$) using GD, perturbed GD, SGD, SVRG, and block coordinate descent. For the SGD method, you may use the mini-batch version.

2) Compare these solvers in terms complexity of hyper-parameter tunning, convergence time, convergence rate (in terms of # outer-loop iterations), and memory requirement

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.

Outline

1. Nonconvex optimization

2. Finding stationary points for structured nonconvex problems

3. Finding stationary points for generic nonconvex problems

4. Finding a local minima

5. Supplements

Application: Matrix factorization

Matrix Factorization: Given a data matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$, find low-rank matrices, $\mathbf{X} \in \mathbb{R}^{N \times k}$ and $\mathbf{Y} \in \mathbb{R}^{M \times k}$, such that $\mathbf{D} \approx \mathbf{X}\mathbf{Y}^T$

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{D} - \mathbf{X}\mathbf{Y}^T\|_F^2$$

Solved using BCD.

- Given $\mathbf{Y}_k$ update $\mathbf{X}_{k+1} := \min_{\mathbf{X}} \ \|\mathbf{D} - \mathbf{X}\mathbf{Y}_k^T\|_F^2$
- Given $\mathbf{X}_{k+1}$ update $\mathbf{Y}_{k+1} := \min_{\mathbf{Y}} \ \|\mathbf{D} - \mathbf{X}_{k+1}\mathbf{Y}^T\|_F^2$
- subproblems are strongly convex: convergence follows from BCD
- a.k.a. Alternating Least Squares
- BCD is basis of many algorithm in recommendation system: non-negative MF, sparse MF, etc.

Application: Sparse dictionary learning

Given a data matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$, find a dictionary $\mathbf{X}\mathbf{Y}^T$, that sparsely represents the data matrix,

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{D} - \mathbf{X}\mathbf{Y}^T\|_F^2 + \lambda \|\mathbf{X}\|_1 \ \text{ s. t. } \|\mathbf{Y}\|_F \leq \beta$$

-

Special cases

Many known algorithms are special cases of block successive upperbound minimization (BSUM)

Difference of convex (DC) programming:

$\arg\min_{\boldsymbol{w}} f(\boldsymbol{w}) = g_1(\boldsymbol{w}) - g_2(\boldsymbol{w})$, where g_1 and g_2 convex

$\boldsymbol{w}^{k+1} = \arg\min_{\boldsymbol{w}} \; g_1(\boldsymbol{w}) - (\nabla \; g_2(\boldsymbol{w}^k)^T(\boldsymbol{w} - \boldsymbol{w}^k)) - g_2(\boldsymbol{w}^k)$

- *Convex Concave Procedure*

Block coordinate descent (BCD):

Select the upperbound in BSUM as the function:
$u_i(\boldsymbol{w}_i, \boldsymbol{w}_{-i}^k) = f(\boldsymbol{w}_i, \boldsymbol{w}_{-i}^k)$

- we recover BCD

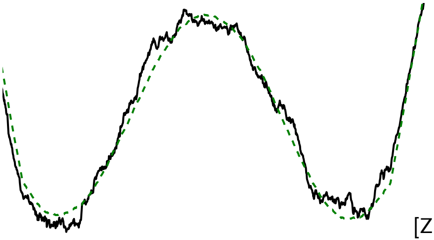
Smooth and optimize

Consider a complicated non-smooth function $f(\boldsymbol{w})$

One may use a smooth approximation $g(\boldsymbol{w}) = \mathbb{E}_{\theta \sim \mathcal{B}(0,R)}[f(\boldsymbol{w} + \theta)]$ for some distribution of θ defined on Borel set $\mathcal{B}$

Random initialization on the obtained smooth function followed by SGD

This approach escapes suboptimal local minima that only exist in the empirical risk (black curve) not the population risk (green curve)



[Zhang-Liang-Charikar, 2017]