The Common-directions Method for Regularized Empirical Risk Minimization

Po-Wei Wang

Department of Computer Science
National Taiwan University

MACHINE LEARNING DEPARTMENT

Machine Learning Department
Carnegie Mellon University

Joint work w/ Ching-Pei Lee (NTU,UW-Madison) & Chih-Jen Lin (NTU)

Talk at SIAM Conference on Optimization, May 2017

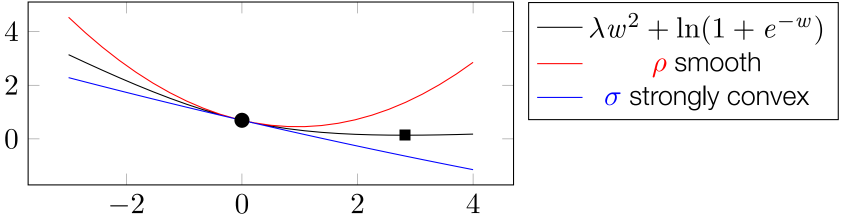
The Settings

Consider the unconstrained optimization problem

$$\underset{\boldsymbol{w} \in \mathbb{R}^n}{\text{minimize}} \quad f(\boldsymbol{w}), \tag{1}$$

where $f(\boldsymbol{w})$ is σ strongly convex and ρ smooth;

$$\frac{1}{2}\boldsymbol{\Delta}^\top(\sigma I)\boldsymbol{\Delta} \leq f(\boldsymbol{w}+\boldsymbol{\Delta}) - f(\boldsymbol{w}) - \nabla f(\boldsymbol{w})^\top \boldsymbol{\Delta} \leq \frac{1}{2}\boldsymbol{\Delta}^\top(\rho I)\boldsymbol{\Delta}.$$



These assumptions give upper and lower bounds for the local second-order Taylor expansion.

Many first- and second-order algorithms generate iterates by solving local approximations.

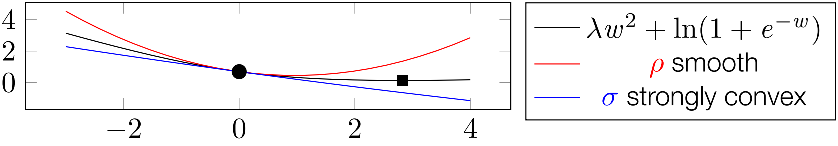
Core Question

Core Question

Many algorithms use past information to refine current approximation.

But current approximate may be biased since Hessian is changing...

So...



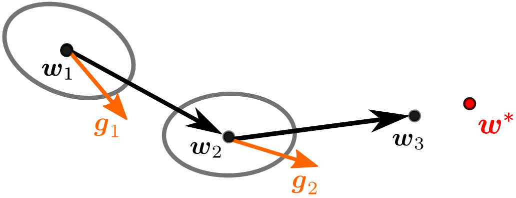
Core Question

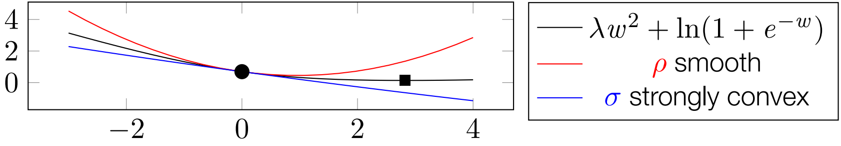
Many algorithms use past information to refine current approximation.

But current approximate may be biased since Hessian is changing...

So...

Should we spend time making more precise local approximation





Core Question

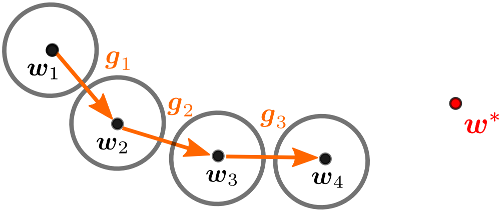
Many algorithms use past information to refine current approximation.

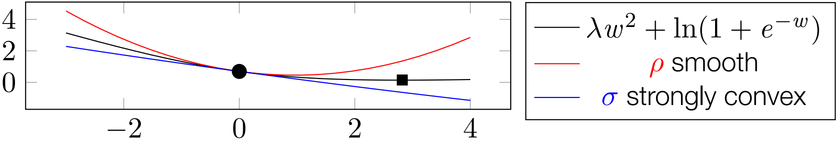
But current approximate may be biased since Hessian is changing...

So...

Should we spend time making more precise local approximation

or should we move quickly because the Hessian is changing?





Core Question

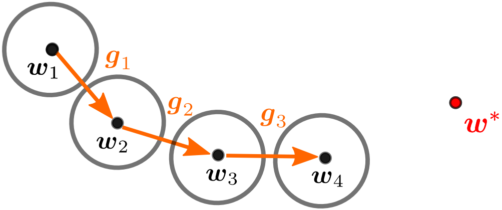
Many algorithms use past information to refine current approximation.

But current approximate may be biased since Hessian is changing...

So...

Should we spend time making more precise local approximation

or should we move quickly because the Hessian is changing?



There is a trade-off in how to use the past information and gradients.

Core Question

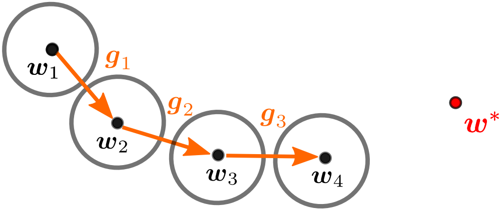
Many algorithms use past information to refine current approximation.

But current approximate may be biased since Hessian is changing...

So...

Should we spend time making more precise local approximation

or should we move quickly because the Hessian is changing?



There is a trade-off in how to use the past information and gradients.

Core Question: How to efficiently reuse past information?

Overview

We present a general framework to reuse previous directions by solving

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t}),$$

in which

$P = \left[\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m\right] \in \mathbb{R}^{n \times m}$ are the basis of past directions.

Each step, we find a approximate solutions in the span of past directions.

Overview

We present a general framework to reuse previous directions by solving

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t}),$$

in which

$P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m] \in \mathbb{R}^{n \times m}$ are the basis of past directions.

Each step, we find a approximate solutions in the span of past directions.

We proved that under certain stopping conditions for subproblems, the method converges globally with the optimal first-order linear rate, and locally with a quadratic rate.

Overview

We present a general framework to reuse previous directions by solving

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t}),$$

in which

$P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m] \in \mathbb{R}^{n \times m}$ are the basis of past directions.

Each step, we find a approximate solutions in the span of past directions.

We proved that under certain stopping conditions for subproblems, the method converges globally with the optimal first-order linear rate, and locally with a quadratic rate.

In the Empirical Risk Minimization problem (ERM), outperforms the state-of-the-art first- and second-order methods in the number of data accesses and is competitive in the running time.

Number of data accesses: # scans through the data, important when data is stored distributedly or cannot fit in the memory

Outline

Background

The Common-directions Method

Theoretical Guarantee

Conclusion

Conjugate Gradient Method

Solve the quadratic problem

$$\underset{\boldsymbol{w}}{\text{minimize}} \ f(\boldsymbol{w}) \equiv \frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w} - \boldsymbol{b}^\top \boldsymbol{w}.$$

Equivalent to our framework by solving

$$\underset{\boldsymbol{t} \in \mathbb{R}^k}{\text{minimize}} \ f(\boldsymbol{w}_k + P\boldsymbol{t}),$$

in which CG gives exact solution $\boldsymbol{t}$ on conjugate basis P,

$$\boldsymbol{p}_k = \boldsymbol{g}_k - \sum_{i<k} \left(\frac{\boldsymbol{g}_k^\top A\boldsymbol{p}_i}{\boldsymbol{p}_i^\top A\boldsymbol{p}_i} \right) \boldsymbol{p}_i, \quad \text{where } \boldsymbol{g}_k = -\nabla f(\boldsymbol{w}_k).$$

Optimal first-order linear rate on positive-definite quadratic problems

Only requires Hessian-vector product $\nabla^2 f(\cdot)\boldsymbol{v}$ in the algorithm

No guarantee on nonlinear cases

Nesterov's Accelerated Method

Constant step-size scheme on alternating sequences $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{s}_k\}$

$$\boldsymbol{s}_k = \boldsymbol{w}_k - \left(\frac{2}{\sqrt{\kappa}+1}\right)(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}),$$

$$\boldsymbol{w}_{k+1} = \boldsymbol{s}_k - \nabla f(\boldsymbol{s}_k),$$

where $\kappa = \dfrac{\rho}{\sigma}$ is the condition number.

In Nesterov's accelerated method, we alternatively use the two directions

$$\boldsymbol{p}_1 = \boldsymbol{w}_k - \boldsymbol{w}_{k-1}, \qquad \boldsymbol{p}_2 = \nabla f(\boldsymbol{s}_k),$$

which is similar to approximately solving the subproblem

$$\underset{\boldsymbol{t} \in \mathbb{R}^2}{\text{minimize}} \ f(\boldsymbol{w}_k + P\boldsymbol{t}),$$

on the above $\boldsymbol{p}_i$, $i = 1, 2$.

Optimal first-order linear rate for the first-order settings

Not strictly decreasing

Quasi-Newton Method

Main idea: approximate the Hessian by past gradients

For example, the BFGS method solves

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}} \ \frac{1}{2}\boldsymbol{w}^\top B_k \boldsymbol{w} + \nabla f(\boldsymbol{w}_k)^\top \boldsymbol{w}$$

by using matrix inversion lemma to maintain the inverse

$$B_k^{-1} = (I - \mu_{k-1}\boldsymbol{u}_{k-1}\boldsymbol{s}_{k-1}^\top) B_{k-1}^{-1} (I - \mu_{k-1}\boldsymbol{u}_{k-1}\boldsymbol{s}_{k-1}^\top) + \mu_{k-1}\boldsymbol{s}_{k-1}\boldsymbol{s}_{k-1}^\top,$$

$$\mu_{k-1} \equiv \frac{1}{\boldsymbol{u}_{k-1}^\top \boldsymbol{s}_{k-1}}, \quad \boldsymbol{s}_{k-1} \equiv \boldsymbol{w}_k - \boldsymbol{w}_{k-1}, \quad \boldsymbol{u}_{k-1} \equiv \nabla f(\boldsymbol{w}_k) - \nabla f(\boldsymbol{w}_{k-1}).$$

If we expand $\boldsymbol{s}_{k-1}$ and $\boldsymbol{u}_{k-1}$ and let $B_0 = \lambda I$, we can see that

$\boldsymbol{w}_{k+1}$ is on the span of past gradients.

Thus, BFGS can be seen as approximately solving the subproblem

$$\underset{\boldsymbol{t} \in \mathbb{R}^k}{\text{minimize}} \ f(\boldsymbol{w}_k + P\boldsymbol{t}),$$

in which

$$\boldsymbol{p}_i = \nabla f(\boldsymbol{w}_i), \quad \forall i = 1, \ldots, k.$$

Summary

Conjugate Gradient Method

- Give exact solution for $\min_{\boldsymbol{t}} f(\boldsymbol{w} + P\boldsymbol{t})$ for quadratic problems
- Solve $\min_{\boldsymbol{t}} f(\boldsymbol{w} + P\boldsymbol{t})$, where P is the conjugate basis

Nesterov's Accelerated Method

- Interpolate between past direction and gradient
- Solve $\min_{\boldsymbol{t}} f(\boldsymbol{w} + P\boldsymbol{t})$, where $P = [\boldsymbol{w}_k - \boldsymbol{w}_{k-1}, \ \nabla f(\boldsymbol{s}_k)]$

Quasi-Newton Method

- Approximate the Hessian with past gradients
- Solve $\min_{\boldsymbol{t}} f(\boldsymbol{w} + P\boldsymbol{t})$, where $P = [\nabla f(\boldsymbol{w}_i)]$, for all $i = 0, \ldots, k$

All above methods involve reusing past information/gradients.

Why not reuse the past gradients directly?

Outline

Background

The Common-directions Method

Theoretical Guarantee

Conclusion

CommDir: Reuse past gradients whenever we can!

In each iteration, we solve the subproblem

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t}),$$

in which $P \in \mathbb{R}^{n \times m}$ is the orthogonal basis of $[\nabla f(\boldsymbol{w}_i)], \quad i = 0, \ldots, k.$

Each step, we find a approximate solutions in the span of past gradients.

Orthogonalized basis P: easier to detect new expansion on $\nabla f(\boldsymbol{w}_k)$.

Alg. (Common-directions Method)

$P = [\nabla f(\boldsymbol{w}_0)/\|\nabla f(\boldsymbol{w}_0)\|]$

For k-th iteration **do**:

- (Approximately) solve subproblem $\min_{\boldsymbol{t}} f(\boldsymbol{w}_k + P\boldsymbol{t})$ to obtain $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + P\boldsymbol{t}$
- Let $\boldsymbol{p} = (I - PP^\top)\nabla f(\boldsymbol{w}_{k+1})$
- **If** $\boldsymbol{p} \neq \boldsymbol{0}$ **then** $P = [P;\ \boldsymbol{p}/\|\boldsymbol{p}\|]$

If the # of iterations is small, then # of variables in the subproblem is also small

Solve the subproblem by (multiple or single iters of) Newton method on $\boldsymbol{t}$ with backtracking line search

If subproblem solved exactly...

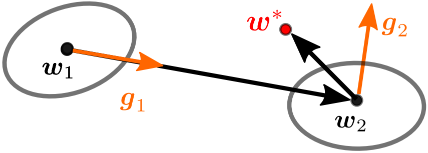
Theorem

When the subproblem $\min_{\boldsymbol{t}} f(\boldsymbol{w}_k + P\boldsymbol{t})$ *is solved exactly,* CommDir *reaches the optimum in n iterations, where $n = \dim(\boldsymbol{w})$.*

Equivalent guarantee to conjugate gradient method on quadratic f, but also works on non-quadratic problems.

Proof.

When subproblem solved exactly, the next gradient direction is orthogonal to all previous directions (otherwise projected gradient is nonzero and the subproblem is not solved exactly). Thus, $\boldsymbol{w}_k + P\boldsymbol{t}$ covers the optimal solution in n iterations. □



Just ideal case. What if the subproblem is solved approximately?

When subproblem solved approximately...

Theorem

Under a proper inner stopping condition,
CommDir *converges globally in an optimal first-order linear rate.*
That is, $f(\boldsymbol{w}_k) - f^* \leq \epsilon$ *in* $O(\sqrt{\kappa}\log(1/\epsilon))$ *iterations, where* $\kappa = \frac{\rho}{\sigma}$.

Subproblem by Newton method on $\boldsymbol{t}$ w/ backtracking line search

Theorem

The line-search procedure terminates in $\lceil \log_\beta(\beta\sigma/(\rho+\lambda)) \rceil$ *steps, where* λ *is the threshold and* β *is the shrinking parameter.*

Even if we only do a single iteration of the inner loop

Theorem

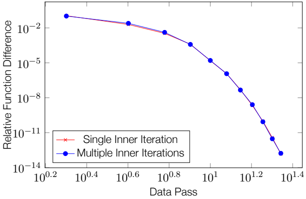
CommDir *with a single inner iteration converges Q-linearly.*
In addition, if the Hessian is Lipschitz continuous,
CommDir *admits local quadratic convergence.*

Optimal guarantees in first- and second-order methods!

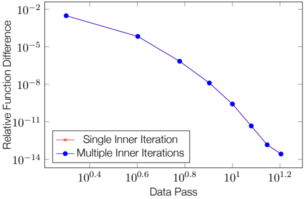
Multiple Inner Iterations v.s. Single Inner Iteration

Experiment suggests that for ERM problems, there is not much difference between using multiple and single inner iterations.

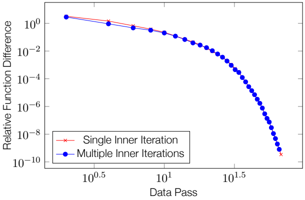
(a) webspam



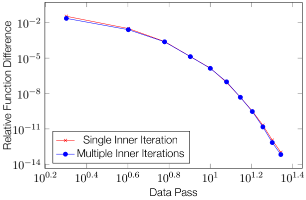
(b) epsilon



(c) url



(d) a9a



Application: CommDir For Empirical Risk Minimization

Despite we only have m variables in the subproblem

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t}),$$

constructing the coefficients for $\boldsymbol{t}$ might be expensive. Need to consider special structure in problems.

Example: Empirical Risk Minimization (SVMs and logistic regression):

$$\underset{\boldsymbol{w}}{\text{minimize}} \; f(\boldsymbol{w}), \text{ where } f(\boldsymbol{w}) \equiv \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{l} \xi(y_i;\; \boldsymbol{w}^\top \boldsymbol{x}_i).$$

The gradient and Hessian have special structure

$$\nabla f(\boldsymbol{w}) = \boldsymbol{w} + X^\top \boldsymbol{v}_{\boldsymbol{w}} \qquad \nabla_{\boldsymbol{t}} f(\boldsymbol{w} + P\boldsymbol{t}) = P^T \boldsymbol{w} + (XP)^\top \boldsymbol{v}_{\boldsymbol{w}}$$

$$\nabla^2 f(\boldsymbol{w}) = I + X^\top D_{\boldsymbol{w}} X \quad \nabla^2_{\boldsymbol{t}} f(\boldsymbol{w} + P\boldsymbol{t}) = I + (XP)^\top D_{\boldsymbol{w}} (XP)$$

Each iteration, we will add at most one direction into P

$$X\left(P, \boldsymbol{p}_{m+1}\right) = \left(XP, X\boldsymbol{p}_{m+1}\right)$$

so we only need to calculate $X\boldsymbol{p}_{m+1}$ to maintain the new XP.

CommDir for ERM Complexity

By proper bookkeeping, the cost per iteration for CommDir is

$$O(\underbrace{lm^2 + mn + \text{\#non-zeros in data}}_{\text{construct gradient and Hessian}} + \underbrace{m^3}_{\text{Newton on subproblem}}),$$

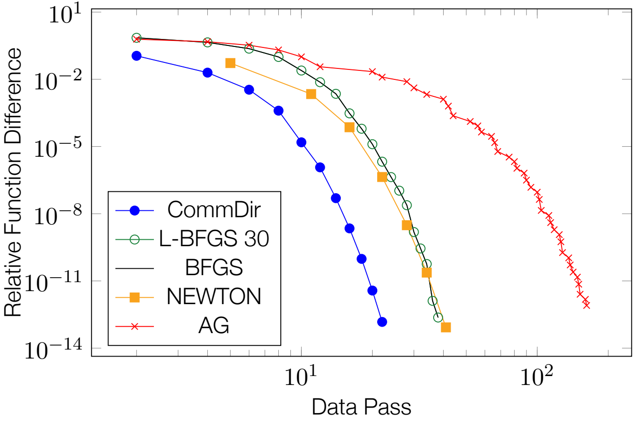
where l is # of data, n is $\dim(\boldsymbol{w})$, and m is the # of stored directions.

Comparable to state-of-the-art methods if m small, and we usually reaches enough precision in 30 iterations ($m \leq 30$).

Experiment: Objective v.s. Data Pass ($C = 10^{-3}$)

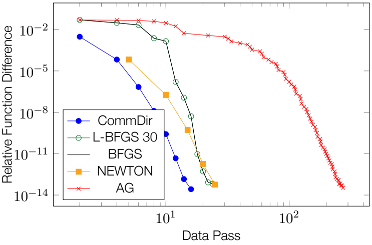
CommDir outperform L-BFGS method w/ 30 past directions, BFGS, truncated Newton method, and Accelerated Gradient method in term of data pass.

(a) webspam

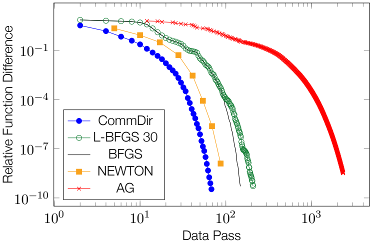


Experiment: Objective v.s. Data Pass ($C = 10^{-3}$)

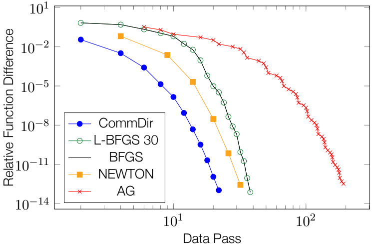
(a) epsilon



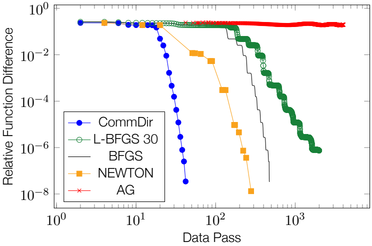
(b) url



(c) a9a



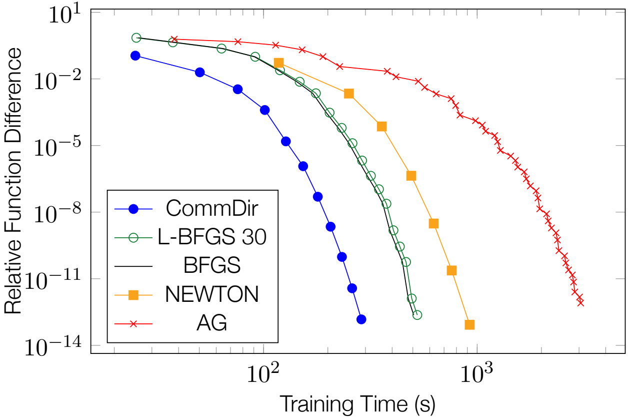
(d) covtype



Experiment: Objective v.s. Time ($C = 10^{-3}$)

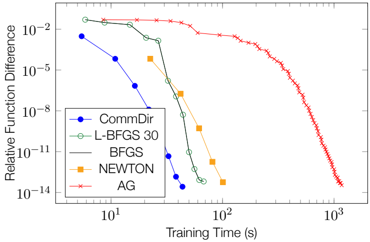
CommDir is also competitive to L-BFGS method w/ 30 past directions, BFGS, truncated Newton method, and Accelerated Gradient method in term of training time.

(a) webspam

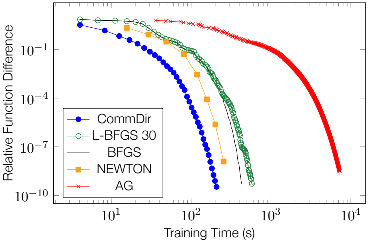


Experiment: Objective v.s. Time ($C = 10^{-3}$)

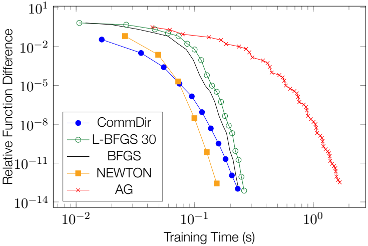
(a) epsilon



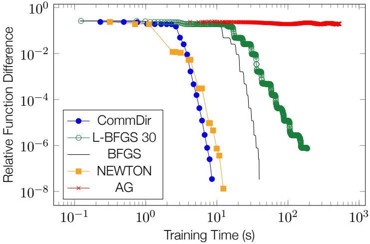
(b) url



(c) a9a



(d) covtype



Summary

We presented the common-directions method, a framework of reusing the past directions.

1. It collects the basis of past directions in P and solve subproblem

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t})$$

2. Under different stopping conditions,
   it admits optimal first-order linear convergence and local quadratic convergence with Liptschitz Hessian.

3. With special structures, e.g. ERM, it can be solved efficiently.

4. Experiments suggest CommDir outperforms BFGS, L-BFGS (with $m = 30$), Nesterov's accelerated gradient method, and truncated Newton method in number of data access, and is competitive in terms of running time.

Now the boring/exciting part: Theoretical Analysis!

Outline

Background

The Common-directions Method

Theoretical Guarantee

Conclusion

Convergence Overview

Common-directions method w/ single inner iteration

- Q-linear convergence
- Local quadratic convergence if Hessian is Lipschitz continuous

Common-directions method w/ multiple inner iterations

- All above properties
- Plus optimal first-order linear rate in $O(\sqrt{\kappa}\log(1/\epsilon))$!

I will just talk about the most interesting part:

Strictly decreasing algorithm with optimal first-order linear rate by reusing past directions.

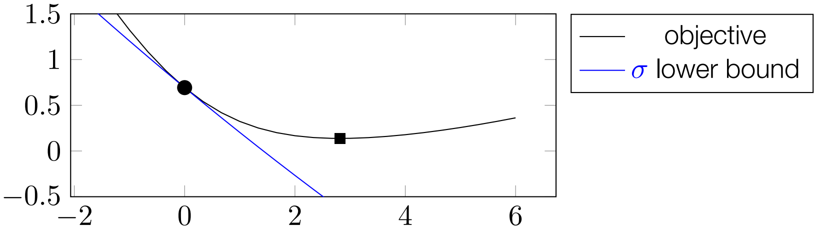
Proof Sketch: Estimation Sequence

Technique from Nesterov's 03 book.

For all $k \geq 0$, recursively define the estimation sequence $\{\phi_k(\boldsymbol{w})\}$ as

$$\phi_{k+1}(\boldsymbol{w}) \equiv (1-\alpha)\phi_k(\boldsymbol{w}) + \alpha \underbrace{\left(f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^\top (\boldsymbol{w} - \boldsymbol{w}_k) + \frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 \right)}_{\text{quadratic lower bound}},$$

$$\text{with} \quad \alpha \equiv \underbrace{\sqrt{\frac{\sigma}{\rho}}}_{\text{rate}} \in (0, 1], \quad \text{and} \quad \phi_0(\boldsymbol{w}) = \underbrace{\frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + f(\boldsymbol{w}_0)}_{\text{initial estimate}}.$$



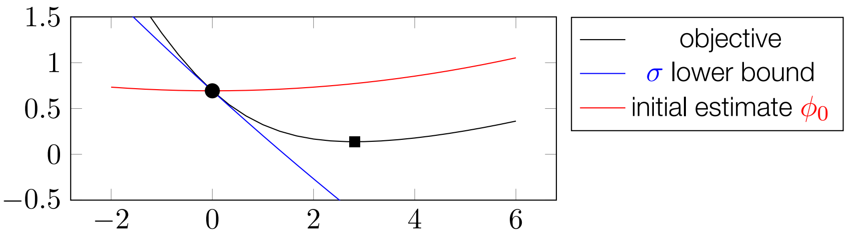
Proof Sketch: Estimation Sequence

Technique from Nesterov's 03 book.

For all $k \geq 0$, recursively define the estimation sequence $\{\phi_k(\boldsymbol{w})\}$ as

$$\phi_{k+1}(\boldsymbol{w}) \equiv (1-\alpha)\phi_k(\boldsymbol{w})+\alpha \underbrace{\left(f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^\top(\boldsymbol{w}-\boldsymbol{w}_k) + \frac{\sigma}{2}\|\boldsymbol{w}-\boldsymbol{w}_k\|^2\right)}_{\text{quadratic lower bound}},$$

$$\text{with } \alpha \equiv \underbrace{\sqrt{\frac{\sigma}{\rho}}}_{\text{rate}} \in (0,1], \text{ and } \phi_0(\boldsymbol{w}) = \underbrace{\frac{\sigma}{2}\|\boldsymbol{w}-\boldsymbol{w}_0\|^2 + f(\boldsymbol{w}_0)}_{\text{initial estimate}}.$$



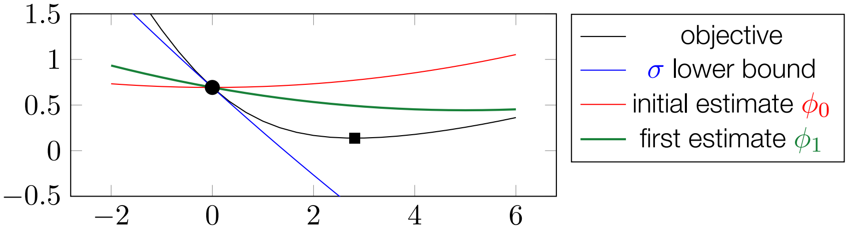
Proof Sketch: Estimation Sequence

Technique from Nesterov's 03 book.

For all $k \geq 0$, recursively define the estimation sequence $\{\phi_k(\boldsymbol{w})\}$ as

$$\phi_{k+1}(\boldsymbol{w}) \equiv (1-\alpha)\phi_k(\boldsymbol{w}) + \alpha \underbrace{\left(f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^\top(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2\right)}_{\text{quadratic lower bound}},$$

$$\text{with } \alpha \equiv \underbrace{\sqrt{\frac{\sigma}{\rho}}}_{\text{rate}} \in (0,1], \quad \text{and } \phi_0(\boldsymbol{w}) = \underbrace{\frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + f(\boldsymbol{w}_0)}_{\text{initial estimate}}.$$



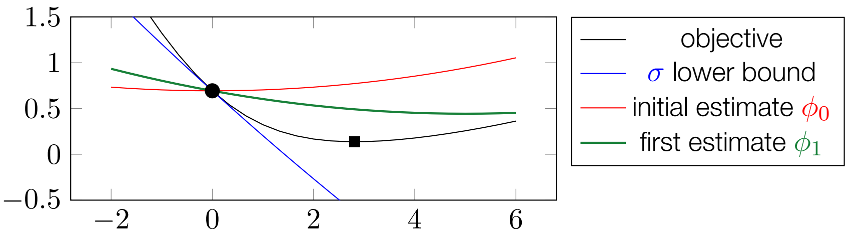
Proof Sketch: Estimation Sequence

Technique from Nesterov's 03 book.

For all $k \geq 0$, recursively define the estimation sequence $\{\phi_k(\boldsymbol{w})\}$ as

$$\phi_{k+1}(\boldsymbol{w}) \equiv (1-\alpha)\phi_k(\boldsymbol{w}) + \alpha \underbrace{\left(f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^\top (\boldsymbol{w} - \boldsymbol{w}_k) + \frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2 \right)}_{\text{quadratic lower bound}},$$

$$\text{with } \alpha \equiv \underbrace{\sqrt{\frac{\sigma}{\rho}}}_{\text{rate}} \in (0, 1], \text{ and } \phi_0(\boldsymbol{w}) = \underbrace{\frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + f(\boldsymbol{w}_0)}_{\text{initial estimate}}.$$



Nesterov's accelerated gradient: generate from estimation sequence.

Proof Sketch: Estimation Sequence

Technique from Nesterov's 03 book.

For all $k \geq 0$, recursively define the estimation sequence $\{\phi_k(\boldsymbol{w})\}$ as

$$\phi_{k+1}(\boldsymbol{w}) \equiv (1-\alpha)\phi_k(\boldsymbol{w}) + \alpha \underbrace{\left(f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^\top (\boldsymbol{w} - \boldsymbol{w}_k) + \frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_k\|^2\right)}_{\text{quadratic lower bound}},$$

$$\text{with } \alpha \equiv \underbrace{\sqrt{\frac{\sigma}{\rho}}}_{\text{rate}} \in (0, 1], \text{ and } \phi_0(\boldsymbol{w}) = \underbrace{\frac{\sigma}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + f(\boldsymbol{w}_0)}_{\text{initial estimate}}.$$

Nesterov's accelerated gradient: generate from estimation sequence.

Key idea:

We do not generate $\boldsymbol{w}_{k+1}$ from the estimation sequence.
Instead, we construct estimation sequence on existing $\boldsymbol{w}_k$ and use it to determine the stopping condition of inner iterations!

Proof Sketch: Inner Stopping Condition

Key idea:

1. Construct estimation sequence ϕ_{k+1} on existing $\boldsymbol{w}_k$
2. Use solution $\boldsymbol{v}_{k+1}$ of estimation sequence $\phi_{k+1}(\cdot)$ for stopping condition for subproblem $\min_{\boldsymbol{t}} f(\boldsymbol{w}_k + P\boldsymbol{t})$.

If the iterate $\boldsymbol{w} = \boldsymbol{w}_k + P\boldsymbol{t}$ on subproblem $\min_{\boldsymbol{t}} f(\boldsymbol{w}_k + P\boldsymbol{t})$ satisfies

$$\text{stopping cond.} \begin{cases} \nabla f(\boldsymbol{w})^\top (\boldsymbol{v}_{k+1} - \boldsymbol{w}) + \frac{\sigma}{2}\|\boldsymbol{v}_{k+1} - \boldsymbol{w}\|^2 \geq 0 & \text{(a)} \\ f(\boldsymbol{w}) \leq f(\boldsymbol{w}_k - \frac{1}{\rho}\nabla f(\boldsymbol{w}_k)) & \text{(b)} \end{cases}$$

$\Longrightarrow$ We are doing better than estimation sequence

$\Longrightarrow$ Optimal first-order rate!

We prove the inner iterations always generate a $\boldsymbol{w}$ satisfying the stopping condition in finite time because we cover the span of $\boldsymbol{v}_{k+1}$.

$\Longrightarrow$ Optimal first-order linear rate.

Reusing previous direction properly is enough for optimal rate!
Interpolation is not required. Strictly decreasing!

Outline

Background

The Common-directions Method

Theoretical Guarantee

Conclusion

Conclusion

In this work, we present the common-directions method, a framework of reusing the past directions.

1. It builds a basis P from past gradients, and solves the subproblem

$$\underset{\boldsymbol{t} \in \mathbb{R}^m}{\text{minimize}} \; f(\boldsymbol{w} + P\boldsymbol{t})$$

2. We got Q-linear convergence and local quadratic convergence (with Lipschitz Hessian) for CommDir with single inner iteration. We got optimal first-order linear convergence for CommDir with multiple inner iterations while being strictly decreasing.

3. We apply CommDir on the empirical risk minimization problems and exploit the structure to make it efficient.

4. Experiments show that it outperforms state-of-the-art first- and second-order optimization methods in the number of data access, and it is also competitive in running time.

Extension: Limited Common-directions Method

What if we limit the length of past directions in the subproblem?
In that case, what kind of directions should we preserve?
What is the convergence guarantee?

Same idea from BFGS to L-BFGS!

We investigate the problem in

C.-P. Lee, P.-W. Wang, W. Chen, and C.-J. Lin.
Limited-memory common-directions method for distributed optimization and its application on empirical risk minimization.
SIAM International Conference on Data Mining, 2017

We found that preserving $\boldsymbol{w}_k - \boldsymbol{w}_{k-1}$ is better than preserving $\nabla f(\boldsymbol{w}_k)$ and proved linear convergence for the scenario.

Thanks for Listening

Please see the full paper at

P.-W. Wang, C.-P. Lee, and C.-J. Lin.
The Common-directions Method for Regularized Empirical Risk Minimization.
Technical report, 2016.
http://www.csie.ntu.edu.tw/~cjlin/papers/nheavy/commdir.pdf

Questions?