Optimization for Machine Learning

John Duchi
Stanford University

London Machine Learning Summer School 2019

Outline

What you should really do

- Go download a copy of Boyd and Vandenberghe's *Convex Optimization*
- Read it, and watch all the lectures on `Youtube`
- Do all the exercises for ee364a/b at Stanford

What is optimization?

Optimization problems

Problem is to

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \;\; x \in X.$$

When is this (efficiently) solvable?

- When things are convex
- *If* we can formulate a numerical problem as minimization of a convex function f over a convex set X, then (roughly) it is solvable

The recipe for all of machine learning

1. Define/find data representation
2. Define a loss measuring performance
3. Minimize the loss

Notation

Optimization notation

- Data will be $A \in \mathbb{R}^{m \times n}$
- Labels/targets $b \in \mathbb{R}^m$
- Optimization variable $x \in \mathbb{R}^n$

m = number of measurements n = dimension

Convex optimization

Convex sets

Definition

A set $C \subset \mathbb{R}^n$ is *convex* if for any $x, y \in C$

$$tx + (1-t)y \in C \quad \text{for all } t \in C$$

Examples

Hyperplane: Let $a \in \mathbb{R}^n$, $b \in \mathbb{R}$,

$$C := \{x \in \mathbb{R}^n : \langle a, x \rangle = b\}.$$

Polyhedron: Let $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, $b \in \mathbb{R}^m$,

$$C := \{x : Ax \leq b\} = \{x \in \mathbb{R}^n : \langle a_i, x \rangle \leq b_i\}$$

Convex functions

A function f is *convex* if its domain $\operatorname{dom} f$ is a convex set and for all $x, y \in \operatorname{dom} f$ we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for all } t \in [0,1].$$

(Define $f(z) = +\infty$ for $z \notin \operatorname{dom} f$)

Example: linear regression

$$\underset{x}{\text{minimize}} \ \frac{1}{2m} \|Ax - b\|_2^2 = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$

Why convex?

Minima of convex functions

Why convex? Let x be a local minimizer of f on the convex set C. Then *global* minmization:

$$f(x) \leq f(y) \text{ for all } y \in C.$$

Subgradients

A vector g is a *subgradient* of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \text{for all } y.$$

Subdifferential

The *subdifferential* (subgradient set) of f at x is

$$\partial f(x) := \{g : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y\}.$$

Subdifferential examples

Let $f(x) = |x| = \max\{x, -x\}$. Then

$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

Optimality and subgradients

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be convex and* $C \subset \mathbb{R}^n$ *be closed convex. Then* $x^\star$ *minimizes* f *over* C *if and only if for some* $g \in \partial f(x^\star)$*,*

$$\langle g, y - x^\star \rangle \geq 0 \text{ for all } y \in C.$$

Beyond convex optimization

Weakly convex functions

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is ρ-*weakly convex* if

$$f(x) + \frac{\rho}{2} \|x - x_0\|_2^2$$

is convex in x for any x_0

Example: smooth functions

Definition

A function f is *L-smooth* if $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|^2$ for all x, y

Remark

Equivalent to $\nabla^2 f(x)$ having all eigenvalues between $-L$ and L, i.e. $\left\|\nabla^2 f(x)\right\|_{\mathrm{op}} \leq L$

Example: compositions

Definition

Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex and $c : \mathbb{R}^n \to \mathbb{R}^m$ be smooth. The function $f(x) = h \circ c(x) = h(c(x))$ is a *composition*.

Theorem

If h is M-Lipschitz and c is L-smooth, then $f = h \circ c$ is $\rho = ML$-weakly convex.

Multclass classification, deep network

Network with sigmoid activations,

$$\sigma(v) = \left[\frac{1}{1+e^{v_j}}\right]_{j=1}^{d},$$

define $z_0 = x$

$$z_i = \sigma(\Theta_i^\top z_{i-1})$$

and loss (at top layer d)

$$\ell(\Theta; z, y) = \log\left(\sum_{i=1}^{k} \exp\left((\theta_i - \theta_y)^\top z\right)\right)$$

Optimization methods

How do we solve optimization problems?

1. Build a “good” but simple local model of f
2. Minimize the model (perhaps regularizing)

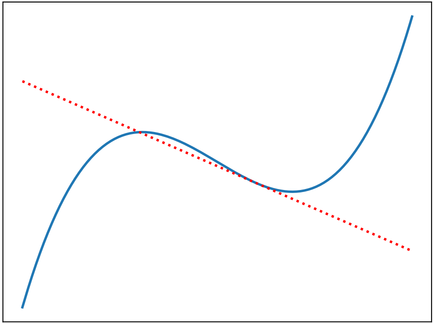
Optimization methods

How do we solve optimization problems?

1. Build a “good” but simple local model of f
2. Minimize the model (perhaps regularizing)

Gradient descent: Taylor (first-order) model

$$f(y) \approx f_x(y) := f(x) + \nabla f(x)^T (y - x)$$



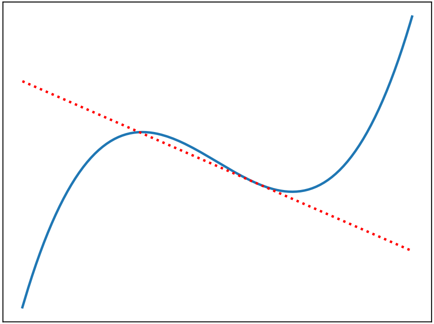
Optimization methods

How do we solve optimization problems?

1. Build a "good" but simple local model of f
2. Minimize the model (perhaps regularizing)

Newton's method: Taylor (second-order) model

$$f(y) \approx f_x(y) := f(x) + \nabla f(x)^T (y - x) + (1/2)(y - x)^T \nabla^2 f(x)(y - x)$$



Composite optimization problems (other model-able structures)

The problem:

$$\underset{x}{\text{minimize}} \;\; f(x) := h(c(x))$$

where

$$h : \mathbb{R}^m \to \mathbb{R} \text{ is convex} \quad \text{and} \quad c : \mathbb{R}^n \to \mathbb{R}^m \text{ is smooth}$$

[Fletcher & Watson 80; Fletcher 82; Burke 85; Wright 87; Lewis & Wright 15; Drusvyatskiy & Lewis 16]

Modeling composite problems

Now we make a *convex* model

$$f(x) = h(c(x))$$

The prox-linear method [Burke, Drusvyatskiy et al.]

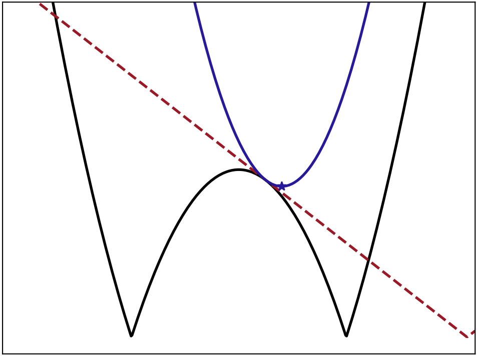
Iteratively (1) form regularized convex model and (2) minimize it

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \end{aligned}$$

The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

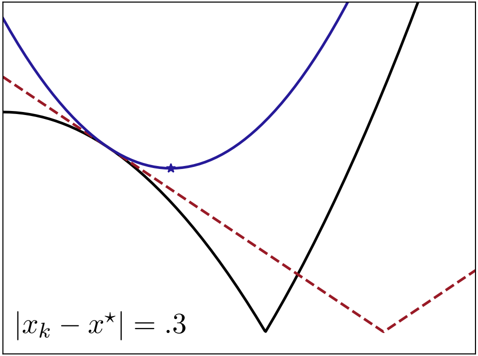
$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$
$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$



The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

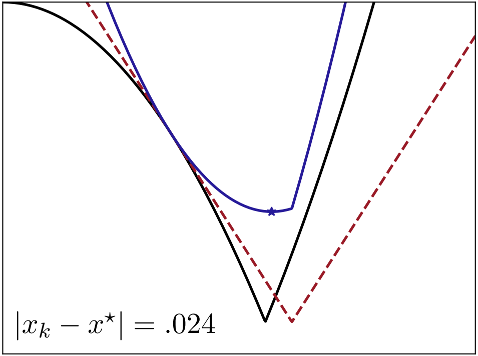
$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \end{aligned}$$



The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

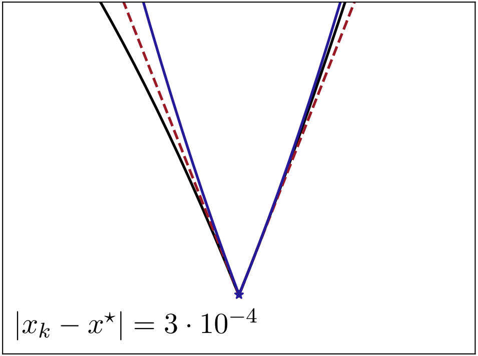
$$
\begin{aligned}
x_{k+1} &= \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k) \right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}
\end{aligned}
$$



The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

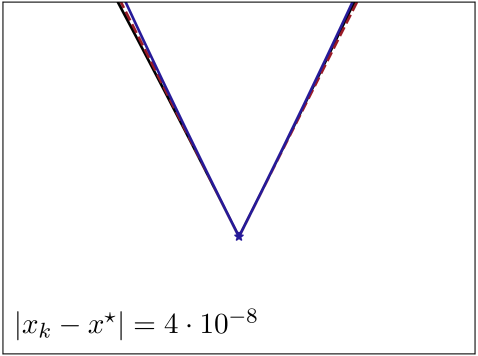
$$
\begin{aligned}
x_{k+1} &= \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}
\end{aligned}
$$



The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

$$
\begin{aligned}
x_{k+1} &= \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}
\end{aligned}
$$



Generic(ish) optimization methods

Iterate

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|^2 \right\}$$

Convex stochastic optimization

Linear regression

- Data: $a_i \in \mathbb{R}^n$, $b_i \in \{\pm 1\}$
- Goal: find x s.t. $a_i^T x \approx b_i$ all i

$$\text{minimize} \;\; f(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2 = \frac{1}{2m} \|Ax - b\|_2^2$$

Challenge: A bit hard when m very very large (even $m = \infty$)

Support vector machines

- Data: $a_i \in \mathbb{R}^n$, $b_i \in \{\pm 1\}$
- Goal: find x s.t. $\mathrm{sign}(a_i^T x) = b_i$ for as many i as possible
- Loss/objective:
$$F(x;(a,b)) = \left[1 - ba^T x\right]_+$$

Stochastic optimization

$$\underset{x}{\text{minimize}} \;\; f(x) = \mathbb{E}[f(x;S)] := \int_{\mathcal{S}} F(x;s)dP(s)$$

where $s \in \mathcal{S}$ is a sample, $S \sim P$ is drawn from population P, instantaneous losses $F(x;S)$

The problem

Problem for now:

$$\underset{x}{\text{minimize}} \ f(x)$$

where f convex, not necessarily differentiable

Gradient method

Consider

$$\underset{x}{\text{minimize}} \;\; f(x)$$

where f convex and continuously differentiable

Gradient method: For some stepsize sequence α_k, iterate

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ &= \underset{x}{\operatorname{argmin}} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\} \end{aligned}$$

Subgradient method

Iterate

$$\text{Choose } \textit{any } g_k \in \partial f(x_k)$$
$$\text{Update } x_{k+1} = x_k - \alpha_k g_k$$

- Not a descent method
- $\alpha_k > 0$ is kth step size

Convergence proof start

A few assumptions to make our lives easier:

- Optimal point: $f^\star = \inf_x f(x) > -\infty$ and there is $x^\star \in \mathbb{R}^n$ with $f(x^\star) = f^\star$
- Lipschitz condition: $\|g\|_2 \leq M$ for all $g \in \partial f(x)$ and all x
- $\|x_1 - x^\star\|_2 \leq R$

(Stronger than needed but whatever)

Convergence proof

Key quantity: distance to optimal point $x^\star$

Convergence proof II

Key step: recursion

Convergence guarantee

Have guarantees

$$\sum_{k=1}^{K} \alpha_k [f(x_k) - f(x^\star)] \le \frac{1}{2} \|x_1 - x^\star\|_2^2 + \sum_{k=1}^{K} \frac{\alpha_k^2}{2} \|g_k\|_2^2$$

or, if $\overline{x}_K = \sum_{k=1}^K \alpha_k x_k / \sum_{k=1}^K \alpha_K$,

$$f(\overline{x}_K) - f(x^\star) \le \frac{R^2 + \frac{1}{2}\sum_{k=1}^K \alpha_k^2 M^2}{\sum_{k=1}^K \alpha_k}$$

Convergence guarantee

For fixed stepsize α and $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$, have

$$f(\overline{x}_K) - f(x^\star) \le \frac{R^2}{\alpha K} + \frac{\alpha}{2} M^2.$$

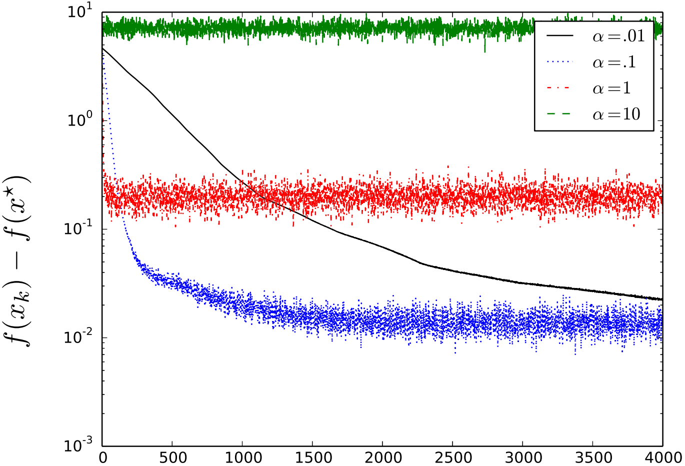
Example: robust regression

$$\text{minimize } f(x) = \frac{1}{m}\|Ax - b\|_1 = \frac{1}{m}\sum_{i=1}^{m} |a_i^T x - b_i|.$$

(Recall: $\partial \|x\|_1 = \mathrm{sign}(x)$, so $\partial f(x) = A^T \mathrm{sign}(Ax - b)$)

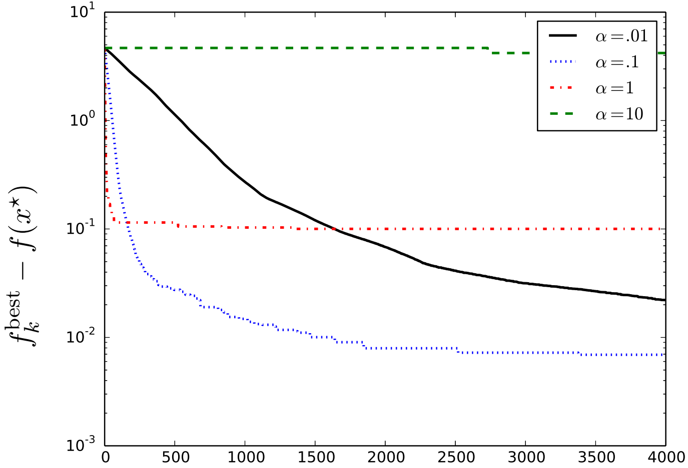
- Perform subgradient descent with fixed stepsize $\alpha \in \{10^{-2}, 10^{-1}, 1, 10\}$.
- Plot $f(x_k) - f^\star$
- Use $f_k^{\mathrm{best}} = \min_{i \le k} f(x_i)$ and plot $f_k^{\mathrm{best}} - f^\star$

Robust regression example



Fixed stepsizes, showing $f(x_k) - f(x^\star)$ for $f(x) = \|Ax - b\|_1$. Here $A \in \mathbb{R}^{100\times 50}$

Robust regression example



Fixed stepsizes, showing $f_k^{\rm best} - f(x^\star)$ for $f(x) = \|Ax - b\|_1$. Here $A \in \mathbb{R}^{100 \times 50}$

Stochastic subgradient methods

Stochastic subgradient: Given function f, a *stochastic* subgradient for a point x is a random vector with

$$\mathbb{E}[g \mid x] \in \partial f(x).$$

Standard example: Expectations. Let S be random variable,

$$f(x) = \mathbb{E}[F(x; S)] = \int F(x; s) dP(s)$$

where $F(\cdot; s)$ is convex. Given x, draw $S \sim P$ and set

$$g = g(x; S) \in \partial F(x; S).$$

(Projected) stochastic subgradient method

Problem:

$$\text{minimize} \;\; f(x) \quad \text{subject to}\, x \in C$$

given access to *stochastic gradients* of f

Method: Iterate with stepsizes $\alpha_k > 0$

- Get stochastic gradient g_k for f at x_k, i.e. $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$
- Update

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k)$$

Motivation and example

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} F(x; S_i)$$

for very large sample $\{S_1, \ldots, S_m\}$.

- True subgradient: take $g_i \in \partial F(x; S_i)$ and

$$g = \frac{1}{m} \sum_{i=1}^{m} g_i$$

- Stochastic subgradient: choose $i \in \{1, \ldots, m\}$ uniformly at random, take $g \in \partial F(x; S_i)$.

Motivation and example

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} F(x; S_i)$$

for very large sample $\{S_1, \ldots, S_m\}$.

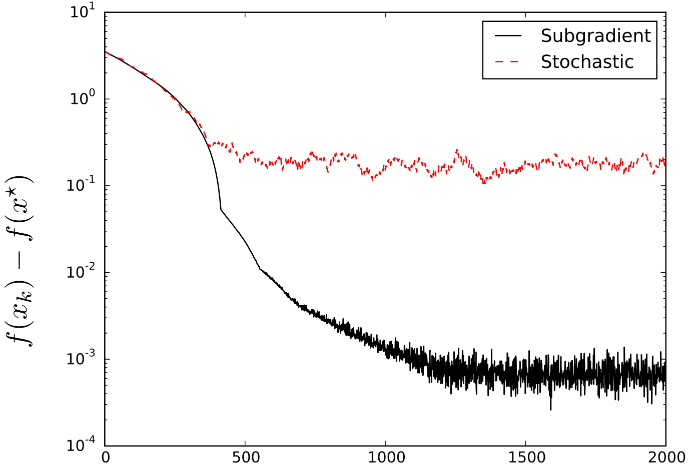
- True subgradient: take $g_i \in \partial F(x; S_i)$ and

$$g = \frac{1}{m}\sum_{i=1}^{m} g_i$$

- Stochastic subgradient: choose $i \in \{1, \ldots, m\}$ uniformly at random, take $g \in \partial F(x; S_i)$.

Example: robust regression

$$f(x) = \frac{1}{m}\|Ax - b\|_1 = \frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle - b_i|.$$



Convergence proof

- Compact set C, so $\|x - y\|_2 \le R$ for all $x, y \in C$
- $\mathbb{E}[\|g\|_2^2] \le M^2$ for stochastic subgradients
- Define error $\xi_k = g_k - f'(x_k)$, where $\mathbb{E}[g_k \mid x_k] = f'(x_k) \in \partial f(x_k)$

Starting point:

$$\|x_{k+1} - x^\star\|_2^2 = \|\pi_C(x_k - \alpha_k g_k) - x^\star\|_2^2 \le \|x_k - \alpha_k g_k - x^\star\|_2^2$$

Convergence proof II

$$\|x_{k+1} - x^\star\|_2^2 \leq \|x_k - x^\star\|_2^2 - 2\alpha_k \langle f'(x_k), x_k - x^\star \rangle + \alpha_k^2 \|g_k\|_2 - 2\alpha_k \langle \xi_k, x_k - x^\star \rangle$$

Convergence of Stochastic Gradient Descent

Final convergence guarantee if C compact and $\|x-y\|_2 \le R$ for $x, y \in C$:

$$\sum_{k=1}^{K} [f(x_k) - f(x^\star)] \le \frac{1}{2\alpha_K} R^2 + \frac{1}{2}\sum_{k=1}^{K} \alpha_k \|g_k\|_2^2 - \sum_{k=1}^{K} \langle \xi_k, x_k - x^\star \rangle.$$

Take Expectations:

Convergence of Stochastic Gradient Descent II

Expected convergence guarantee: If $\alpha_k = R/M\sqrt{k}$ and $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$,

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leq \frac{3}{2}\frac{RM}{\sqrt{K}}.$$

Model-based methods

Generic(ish) optimization methods

Iterate

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

Model-based stochastic optimization

Iterate:

- Sample $S_k \overset{\text{iid}}{\sim} P$
- Update by minimizing model

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ F_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

Models in stochastic convex optimization

Conditions on our models

i. Convex model:

$$y \mapsto F_x(y; s) \quad \text{is convex}$$

ii. Lower bound:

$$F_x(y; s) \le F(y; s)$$

iii. Local correctness:

$$F_x(x; s) = F(x; s) \quad \text{and} \quad \partial F_x(x; s) \subset \partial F(x; s)$$

[D. & Ruan 17; Davis & Drusvyatskiy 18]

Example models

Example: linear regression

Example: truncated model and update

Convergence guarantees

Idea: Always as good convergence as subgradient method

Theorem (Davis & Drusvyatskiy 18, Asi & D. 19)

Suppose that models satisfy conditions and $\mathbb{E}[\|g\|^2] \le M^2$ *for stochastic gradients. Then*

$$\mathbb{E}[f(\overline{x}_k)] - f(x^\star) \le \frac{\|x_1 - x^\star\|_2^2}{2\sum_{i=1}^k \alpha_i} + \frac{\sum_{i=1}^k \alpha_i^2 M^2}{\sum_{i=1}^k \alpha_i}.$$

Proof of convergence

Starting point: Optimality of iterate. For $g \in \partial f(x_{k+1}; S_k)$,

$$\left\langle g + \frac{1}{\alpha_k}(x_{k+1} - x_k), x - x_{k+1} \right\rangle \geq 0 \text{ all } x \in X.$$

Proof of convergence II

Iterate recursion:

$$\begin{aligned}\frac{1}{2}\left\|x_{k+1}-x\right\|_2^2 \le \frac{1}{2}\left\|x_k-x\right\|_2^2 &+ \alpha_k[f(x;S_k)-f(x_k;S_k)] \\ &- \alpha_k\langle f'(x_k;S_k), x_{k+1}-x_k\rangle - \frac{1}{2}\left\|x_k-x_{k+1}\right\|_2^2\end{aligned}$$

Stability guarantees (convex)

Use full stochastic-proximal method,

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$

Theorem (Asi & D. 18)

Assume $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ *is non-empty and* $\mathbb{E}[\|f'(x^\star; S)\|^2] \leq \sigma^2$*. Then*

$$\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \leq \operatorname{dist}(x_0, \mathcal{X}^\star)^2 + \sigma^2 \sum_{i=1}^{k} \alpha_i^2$$

Stability guarantees (convex)

Use full stochastic-proximal method,

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$

Theorem (Asi & D. 18)

Assume $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ *is non-empty and* $\mathbb{E}[\|f'(x^\star; S)\|^2] \le \sigma^2$*.*
Then

$$\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \le \operatorname{dist}(x_0, \mathcal{X}^\star)^2 + \sigma^2 \sum_{i=1}^{k} \alpha_i^2$$

Theorem (Asi & D. 18)

Under the same assumptions,

$$\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \textit{and} \quad \operatorname{dist}(x_k, \mathcal{X}^\star) \overset{a.s.}{\to} 0.$$

Stability guarantees (convex)

Use any model with $f_x(y;s) \geq \inf_z f(z;s)$ (i.e. good lower bound)

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x;S_k) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|^2 \right\}.$$

Theorem (Asi & D. 19)

Assume $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ *is non-empty and there exists* $p < \infty$ *such that*

$$\mathbb{E}[\left\| f'(x;S) \right\|^2] \leq C(1 + \operatorname{dist}(x, \mathcal{X}^\star)^p).$$

Then

$$\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \textit{and} \quad \operatorname{dist}(x_k, \mathcal{X}^\star) \stackrel{a.s.}{\to} 0.$$

Classical asymptotic analysis

Theorem (Polyak & Juditsky 92)

Let F be convex and strongly convex in a neighborhood of $x^\star$, and assume that $f(x;S)$ are globally smooth. For x_k generated by stochastic gradient method,

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \nabla^2 F(x^\star)^{-1}\right).$$

New asymptotic analysis (convex case)

Theorem (Asi & D. 18)

Let F be convex and strongly convex in a neighborhood of $x^\star$, and assume that $f(x; S)$ are smooth near $x^\star$. Then if x_k remain bounded and the models $f_{x_k}(\cdot; S_k)$ satisfy our conditions,

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \nabla^2 F(x^\star)^{-1}\right).$$

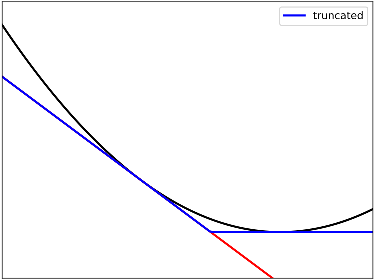
New asymptotic analysis (convex case)

Theorem (Asi & D. 18)

Let F be convex and strongly convex in a neighborhood of $x^\star$, and assume that $f(x;S)$ are smooth near $x^\star$. Then if x_k remain bounded and the models $f_{x_k}(\cdot;S_k)$ satisfy our conditions,

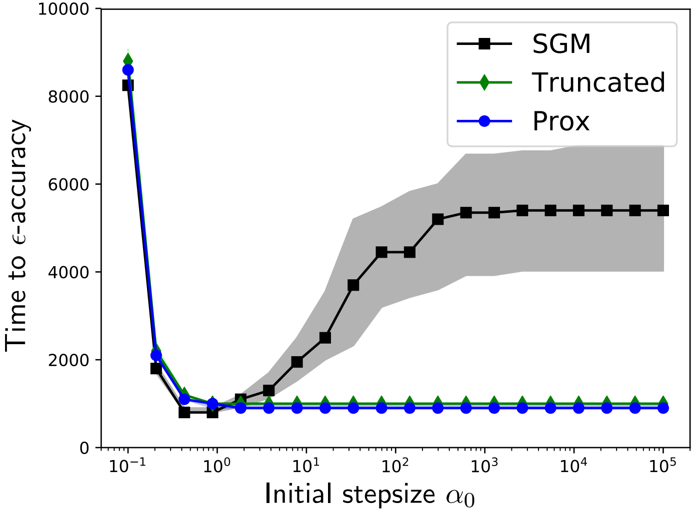
$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}(x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1}\,\mathrm{Cov}(\nabla f(x^\star;S))\nabla^2 F(x^\star)^{-1}\right).$$

- Optimal by local minimax theorem [Hájek 72; Le Cam 73; D. & Ruan 19]
- Key insight: subgradients of $f_{x_k}(\cdot;S_k)$ close to $\nabla f(x_k;S_k)$



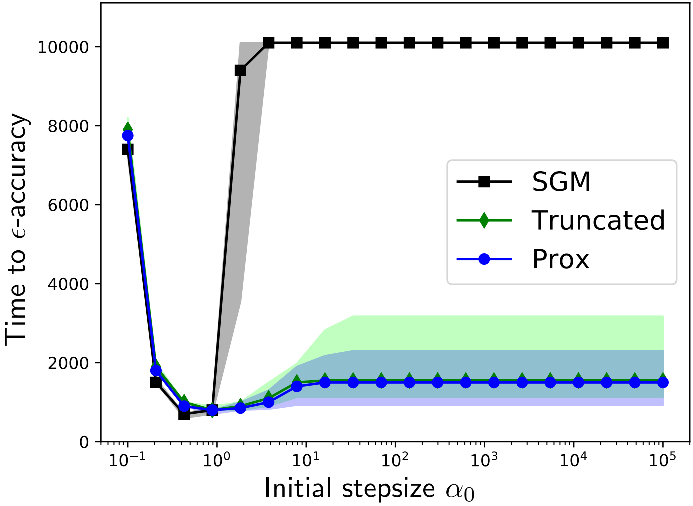
Multiclass hinge loss: no noise

$$f(x;(a,l)) = \max_{i \neq l} \left[1 + \langle a, x_i - x_l \rangle\right]_+$$



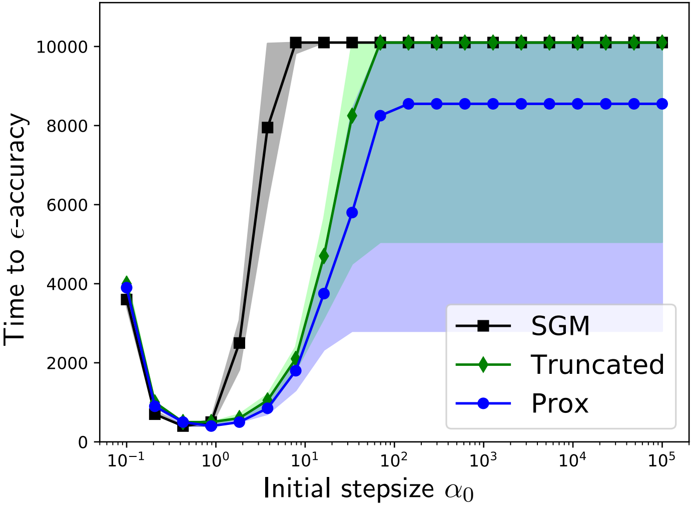
Multiclass hinge loss: small label flipping

$$f(x; (a, l)) = \max_{i \neq l} [1 + \langle a, x_i - x_l \rangle]_+$$



Multiclass hinge loss: substantial label flipping

$$f(x;(a,l)) = \max_{i \neq l} \left[1 + \langle a, x_i - x_l \rangle\right]_+$$



Beyond convex stochastic optimization

Weakly convex optimization

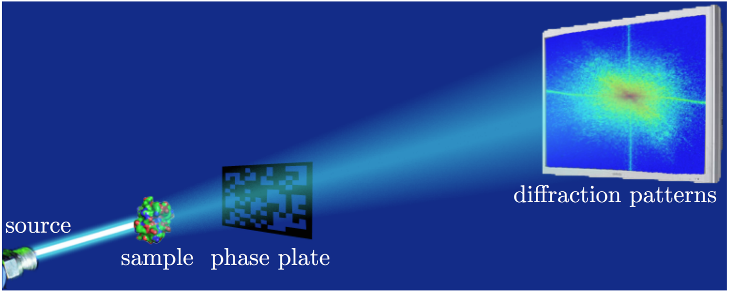
Recall f is *ρ-weakly convex* if

$$f(x) + \frac{\rho}{2} \|x - x_0\|_2^2$$

is convex for any x_0

Motivating problems

(Robust) Phase retrieval



[Candès, Li, Soltanolkotabi 15]

Observations (usually)

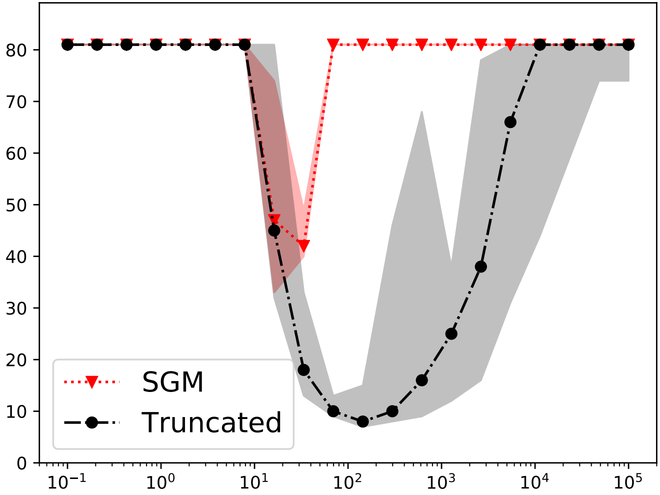
$$b_i = \langle a_i, x^\star \rangle^2$$

yield objective

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i|$$

Matrix completion

$$f(x,y) = \sum_{i,j\in\Omega} |\langle x_i, y_j\rangle - B_{ij}|$$



Convergence for weakly convex problems

- **Issue:** No longer can get to minima
- **More issues:** What are stationary points? Are there subgradients?
- **Even more issues:** Even in the convex case, getting to zero (sub)gradient can be hard

Subgradients for weakly convex functions

Definition

For a ρ-weakly convex f, the subdifferential is

$$\partial f(x) := \left\{ g \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle g, y - x \rangle - \frac{\rho}{2} \|y - x\|^2 \text{ all } y \right\}$$

Equivalent version:

$$\partial f(x) := \partial_y \left\{ f(y) + \frac{\rho}{2} \|y - x\|^2 \right\} \Big|_{y=x}$$

Example: if $f(x) = h(c(x))$ for h convex, c smooth,

$$\partial f(x) = \nabla c(x)^T \partial h(c(x))$$

Example subgradients

- Phase retrieval term: $f(x) = |\langle a, x\rangle^2 - b|$ has

$$\partial f(x) = \operatorname{sign}(\langle a, x\rangle^2 - b) a a^T x$$

  where $\operatorname{sign}(0) = [-1, 1]$

- Matrix completion term: $f(x) = |\langle x, y\rangle - b|$ has

$$\partial f(x, y) = \begin{bmatrix} \operatorname{sign}(\langle x, y\rangle - b) y \\ \operatorname{sign}(\langle x, y\rangle - b) x \end{bmatrix}$$

Proximal regularization

Definition

For $f : \mathbb{R}^n \to \mathbb{R}$ a ρ-weakly convex function, the *proximal regularization* or *Moreau-envelope* of f is

$$f_\lambda(x) := \inf_y \left\{ f(y) + \frac{\lambda}{2} \|y - x\|^2 \right\}$$

The proximal operator is

$$\mathsf{prox}_f^\lambda(y) := \operatorname*{argmin}_y \left\{ f(y) + \frac{\lambda}{2} \|y - x\|^2 \right\}$$

Proposition

If $\lambda > \rho$, then

$$\nabla f_\lambda(x) = \lambda(x - \mathsf{prox}_f^\lambda(x))$$

Proximal regularization: the target of convergence

Nice properties:

- Improvement in objective: $f(\mathsf{prox}_f^\lambda(x)) \leq f(x)$
- Near stationarity: if

$$\left\|\mathsf{prox}_f^\lambda(x) - x\right\| \leq \epsilon \quad \text{then} \quad \left\|\partial f(\mathsf{prox}_f^\lambda(x))\right\| \leq \lambda\epsilon$$

(i.e. ϵ-close to ϵ-stationary point)

Proximal regularization: the target of convergence

Nice properties:

- Improvement in objective: $f(\mathsf{prox}_f^\lambda(x)) \leq f(x)$
- Near stationarity: if

$$\left\|\mathsf{prox}_f^\lambda(x) - x\right\| \leq \epsilon \quad \text{then} \quad \left\|\partial f(\mathsf{prox}_f^\lambda(x))\right\| \leq \lambda\epsilon$$

(i.e. ϵ-close to ϵ-stationary point)

Example: $f(x) = |x|$ has

$$f_\lambda(x) = \begin{cases} \frac{\lambda}{2}x^2 & \text{if } |x| \leq \frac{1}{\lambda} \\ |x| - \frac{1}{2\lambda} & \text{if } |x| > \frac{1}{\lambda} \end{cases}$$

So game is to find points with $\mathsf{prox}_f^\lambda(x) \approx x$

Model-based methods

Models in stochastic convex optimization

Conditions on our models

i. Convex model:

$$y \mapsto F_x(y; s) \quad \text{is convex}$$

ii. Lower bound:

$$F_x(y; s) \le F(y; s) + \frac{\rho(s)}{2} \|y - x\|^2$$

iii. Local correctness:

$$F_x(x; s) = F(x; s) \quad \text{and} \quad \partial F_x(x; s) \subset \partial F(x; s)$$

[D. & Ruan 17; Davis & Drusvyatskiy 18; Asi & D. 19]

Example: matrix completion

Convergence guarantees

Method: Iterate

$$\text{Draw } S_k \stackrel{\text{iid}}{\sim} P$$

$$\text{Update } x_{k+1} = \operatorname*{argmin}_x \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

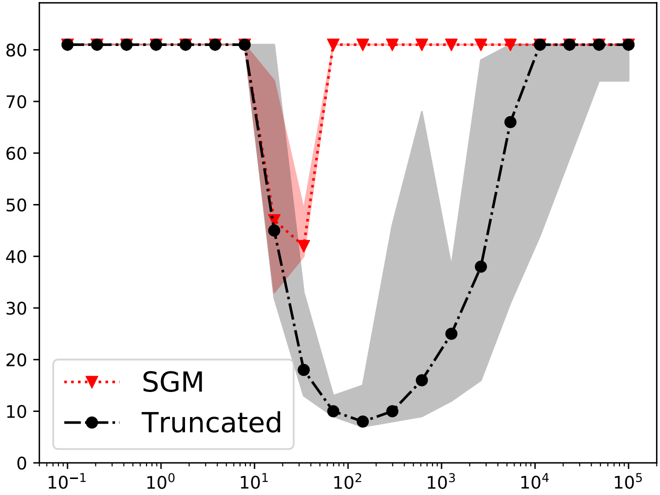
Theorem (Davis & Drusvyatskiy 18)

Assume that $\mathbb{E}[\|f'(x;S)\|^2] \le M^2$. *Then*

$$\mathbb{E}\left[\sum_{i=1}^{k} \alpha_i \|\nabla f_\lambda(x_i)\|^2 \right] \le f(x_1) - f(x^\star) + \frac{\lambda}{2} \sum_{i=1}^{k} \alpha_i^2 M^2.$$

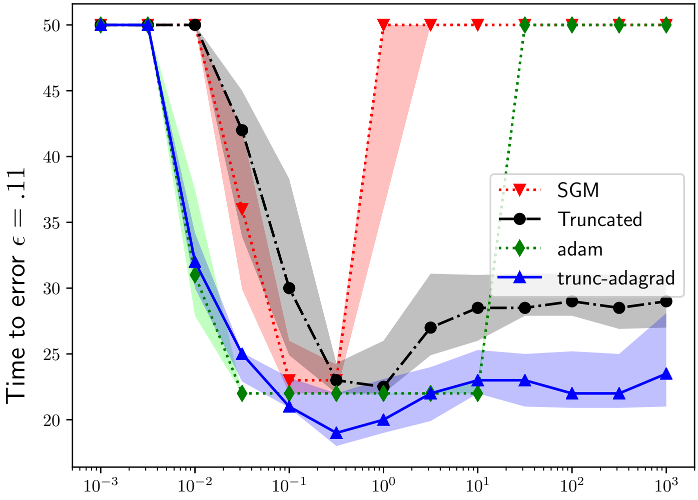
Matrix completion without noise

$$F(x, y) = \sum_{i,j \in \Omega} |\langle x_i, y_j \rangle - M_{ij}|$$



Deep learning experiments

CIFAR 10 Dataset: 10 class image classification



Sharp convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m} \sum_{i=1}^m \left[1 - a_i^T x\right]_+$
- Projection onto intersections: $F(x) = \frac{1}{m} \sum_{i=1}^m \operatorname{dist}(x, C_i)$

Sharp convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m} \sum_{i=1}^m \left[1 - a_i^T x\right]_+$
- Projection onto intersections: $F(x) = \frac{1}{m} \sum_{i=1}^m \operatorname{dist}(x, C_i)$

Theorem (Asi & D. 19)

Let F have sharp growth and be easy. If F is convex,

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, X^\star)^2] \leq \max\left\{\exp(-ck), \exp\left(-c\sum_{i=1}^k \alpha_i\right)\right\} \operatorname{dist}(x_1, X^\star)^2.$$

Sharp weakly convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Phase retrieval $F(x) = \frac{1}{m} \left\| (Ax)^2 - (Ax^\star)^2 \right\|_1$
- Blind deconvolution [Charisopoulos et al. 19]

Sharp weakly convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Phase retrieval $F(x) = \frac{1}{m} \left\| (Ax)^2 - (Ax^\star)^2 \right\|_1$
- Blind deconvolution [Charisopoulos et al. 19]

Theorem (Asi & D. 19)

Let F have sharp growth and be easy. There exists $c \in (0, 1)$ such that on the event $x_k \to X^\star$,

$$\limsup_k \frac{\operatorname{dist}(x_k, X^\star)}{(1-c)^k} < \infty.$$

Conclusions

- Perhaps blind application of stochastic gradient methods is not the right answer
- Care and better modeling can yield improved performance
- Computational efficiency important in model choice

Partial references

- H. Asi and J. C. Duchi. The importance of better models in stochastic optimization. *arXiv:1903.08619 [math.OC]*, 2019
- H. Asi and J. C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, To Appear, 2019. URL https://arXiv.org/abs/1810.05633
- D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming, Series B*, 129:163–195, 2011
- P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004
- J. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33:260–279, 1985
- J. Burke and M. Ferris. A Gauss-Newton method for convex composite optimization. *Mathematical Programming*, 71:179–194, 1995
- D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019
- D. Drusvyatskiy and A. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018
- D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming, Series A*, To Appear, 2018

Partial references

- J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval.
  Information and Inference, iay015, 2018
- J. C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems.
  SIAM Journal on Optimization, 28(4):3229–3259, 2018
- R. Fletcher. A model algorithm for composite nondifferentiable optimization problems.
  Mathematical Programming Study, 17:67–76, 1982
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming.
  SIAM Journal on Optimization, 19(4):1574–1609, 2009
- N. Parikh and S. Boyd. Proximal algorithms.
  Foundations and Trends in Optimization, 1(3):123–231, 2013
- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent.
  In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003