

Optimal first-order minimization methods

with applications to image reconstruction and ML



Donghwan Kim & Jeffrey A. Fessler

EECS Dept., BME Dept., Dept. of Radiology
University of Michigan

<http://web.eecs.umich.edu/~fessler>



MSU CMSE Seminar

2016-09-12

- Research support from GE Healthcare
- Supported in part by NIH grant U01 EB018753
- Equipment support from Intel Corporation

Lower-dose X-ray CT image reconstruction



Thin-slice FBP
Seconds

ASIR
A bit longer

Statistical
Much longer

Image reconstruction as an optimization problem:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \succeq \mathbf{0}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_{\mathbf{W}}^2 + R(\mathbf{x}),$$

\mathbf{y} data, \mathbf{A} system model, \mathbf{W} statistics, $R(\mathbf{x})$ regularizer.

(Same sinogram, so all at same dose.)

Motivation

Problem setting

Existing algorithms

- Gradient descent

- Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

- Logistic regression for machine learning

- CT image reconstruction

- Further acceleration using OS

Generalizing OGM

Summary / future work

Motivation

Problem setting

Existing algorithms

- Gradient descent

- Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

- Logistic regression for machine learning

- CT image reconstruction

- Further acceleration using OS

Generalizing OGM

Summary / future work

$$\hat{\mathbf{x}} \in \arg \min_{\mathbf{x}} f(\mathbf{x})$$

- ▶ Unconstrained
- ▶ Large-scale (Hessian $\nabla^2 f$ too big to store and/or undefined)
 - ▶ image reconstruction / inverse problems
 - ▶ big-data / machine learning
 - ▶ ...
- ▶ Cost function assumptions (throughout)
 - ▶ $f : \mathbb{R}^M \mapsto \mathbb{R}$
 - ▶ convex (need not be strictly convex)
 - ▶ non-empty set of global minimizers:

$$\hat{\mathbf{x}} \in \mathcal{X}^* = \{\mathbf{x}_* \in \mathbb{R}^M : f(\mathbf{x}_*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^M\}$$

- ▶ smooth (differentiable with L -Lipschitz gradient)

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{z})\|_2 \leq L \|\mathbf{x} - \mathbf{z}\|_2, \quad \forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^M$$

Example: Fair potential function

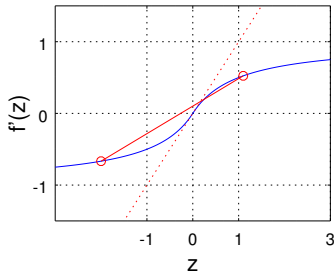
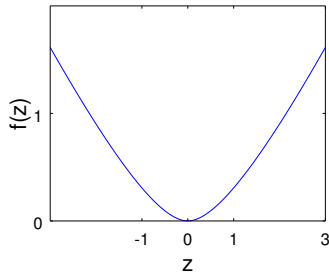
Fair's potential function [1]
(similar to Huber function
and hyperbola):

$$\psi(z) = \delta^2 [|z/\delta| - \log(1 + |z/\delta|)]$$

$$\dot{\psi}(z) = \frac{z}{1 + |z/\delta|}$$

$$\ddot{\psi}(z) = \frac{1}{(1 + |z/\delta|)^2} \leq 1.$$

Thus $L = 1$.



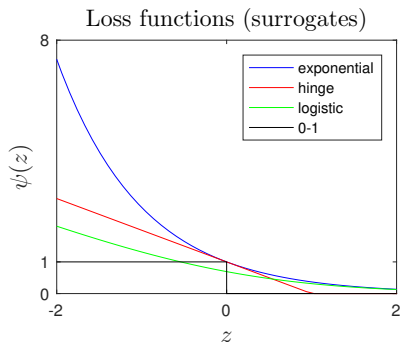
Example: Machine learning

To learn weights \mathbf{x} of binary classifier given feature vectors $\{\mathbf{v}_i\}$ and labels $\{y_i = \pm 1\}$:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} f(\mathbf{x}), \quad f(\mathbf{x}) = \sum_i \psi(y_i \langle \mathbf{x}, \mathbf{v}_i \rangle).$$

loss functions $\psi(z)$

- ▶ 0-1: $\mathbb{I}_{\{z \leq 0\}}$
- ▶ exponential: $\exp(-z)$
- ▶ logistic: $\log(1 + \exp(-z))$
- ▶ hinge: $\max\{0, 1 - z\}$



Which of these ψ fit our conditions?

Motivation

Problem setting

Existing algorithms

Gradient descent

Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

Logistic regression for machine learning

CT image reconstruction

Further acceleration using OS

Generalizing OGM

Summary / future work

- ▶ Problem:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} f(\mathbf{x}).$$

- ▶ Initial guess \mathbf{x}_0 .
- ▶ Simple recursive iteration:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \nabla f(\mathbf{x}_n).$$

- ▶ Step size $1/L$ ensures monotonic descent of f .
- ▶ Telescoping (for intuition, not implementation):

$$\mathbf{x}_{n+1} = \mathbf{x}_0 - \frac{1}{L} \sum_{k=0}^n \nabla f(\mathbf{x}_k).$$

- ▶ Classic $O(1/n)$ convergence rate of cost function descent:

$$\underbrace{f(\mathbf{x}_n) - f(\mathbf{x}_*)}_{\text{inaccuracy}} \leq \frac{L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{2n}.$$

- ▶ Drori & Teboulle (2014) derive tight inaccuracy bound:

$$f(\mathbf{x}_n) - f(\mathbf{x}_*) \leq \frac{L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{4n + 2}.$$

- ▶ They construct a Huber-like function f for which GD achieves that bound \implies case closed for GD with step size $1/L$.
- ▶ $O(1/n)$ rate is undesirably slow.

- ▶ GD with general step size h :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{h}{L} \nabla f(\mathbf{x}_n).$$

- ▶ Classical monotone descent result:
 $h \in (0, 2) \implies f(\mathbf{x}_{n+1}) < f(\mathbf{x}_n)$ when \mathbf{x}_n is not a minimizer.
- ▶ What is best h ?
- ▶ If f is quadratic, then *asymptotic* best choice is:

$$h_* = \frac{2L}{\lambda_{\max}(\nabla^2 f) + \lambda_{\min}(\nabla^2 f)}.$$

- ▶ GD with general step size h :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{h}{L} \nabla f(\mathbf{x}_n).$$

- ▶ More generally, Taylor et al. [3] recently conjectured:

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) \leq \frac{L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{2} \max \left\{ \frac{1}{2Nh + 1}, (1 - h)^{2N} \right\}.$$

- ▶ Proof for $0 < h \leq 1$ by Drori and Teboulle [2]
- ▶ Upper bounds achieved by Huber-like function and quadratic function $f(x) = (L/2)x^2$ respectively.
- ▶ Best h depends on N !
(For $N = 1$, $h_* = 1.5$; for $N = 100$, $h_* = 1.9705$.)
- ▶ Must select N in advance?
- ▶ Still $O(1/N)$...

- ▶ Quest for accelerated convergence.
- ▶ Heavy ball iteration (Polyak, 1987):

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{\alpha}{L} \nabla f(\mathbf{x}_n) + \underbrace{\beta (\mathbf{x}_n - \mathbf{x}_{n-1})}_{\text{momentum!}}$$

(recursive form
to implement)

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n \underbrace{\alpha \beta^{n-k}}_{\text{coefficients}} \nabla f(\mathbf{x}_k)$$

(summation form
to analyze)

- ▶ How to choose α and β ?
- ▶ How to optimize coefficients more generally?

- ▶ General “first-order” (GFO) method:

$$\mathbf{x}_{n+1} = \text{function}(\mathbf{x}_0, f(\mathbf{x}_0), \nabla f(\mathbf{x}_0), \dots, f(\mathbf{x}_n), \nabla f(\mathbf{x}_n)).$$

- ▶ First-order (FO) methods with fixed step-size coefficients:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n h_{n+1,k} \nabla f(\mathbf{x}_k).$$

Primary goals:

- ▶ Analyze convergence rate of FO for any given $\{h_{n,k}\}$
- ▶ Optimize step-size coefficients $\{h_{n,k}\}$
 - ▶ fast convergence
 - ▶ efficient recursive implementation
 - ▶ universal (design *prior* to iterating, independent of L)

Barzilai & Borwein, 1988:

$$\mathbf{g}^{(n)} \triangleq \nabla f(\mathbf{x}_n)$$

$$\alpha_n = \frac{\|\mathbf{x}_n - \mathbf{x}_{n-1}\|_2^2}{\langle \mathbf{x}_n - \mathbf{x}_{n-1}, \mathbf{g}^{(n)} - \mathbf{g}^{(n-1)} \rangle}$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha_n \nabla f(\mathbf{x}_n).$$

- ▶ In “general” first-order (GFO) class, but
- ▶ not in class FO with fixed step-size coefficients.
- ▶ Likewise for methods like
 - ▶ steepest descent (with line search),
 - ▶ conjugate gradient,
 - ▶ quasi-Newton ...

Nesterov's fast gradient method (FGM1)

Nesterov (1983) iteration: Initialize: $t_0 = 1$, $\mathbf{z}_0 = \mathbf{x}_0$

$$\mathbf{z}_{n+1} = \mathbf{x}_n - \frac{1}{L} \nabla f(\mathbf{x}_n) \quad (\text{usual GD update})$$

$$t_{n+1} = \frac{1}{2} \left(1 + \sqrt{1 + 4t_n^2} \right) \quad (\text{magic momentum factors})$$

$$\mathbf{x}_{n+1} = \mathbf{z}_{n+1} + \frac{t_n - 1}{t_{n+1}} (\mathbf{z}_{n+1} - \mathbf{z}_n) \quad (\text{update with momentum}) .$$

Reverts to GD if $t_n = 1, \forall n$.

FGM1 is in class FO:
$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n h_{n+1,k} \nabla f(\mathbf{x}_k)$$

$$h_{n+1,k} = \begin{cases} \frac{t_n - 1}{t_{n+1}} h_{n,k}, & k = 0, \dots, n-2 \\ \frac{t_n - 1}{t_{n+1}} (h_{n,n-1} - 1), & k = n-1 \\ 1 + \frac{t_n - 1}{t_{n+1}}, & k = n. \end{cases}$$

[1	0	0	0	0	0
	0	1.25	0	0	0	0
	0	0.10	1.40	0	0	0
	0	0.05	0.20	1.50	0	0
	0	0.03	0.11	0.29	1.57	0
	0	0.02	0.07	0.18	0.36	1.62
]						

Shown by Nesterov to be $O(1/n^2)$ for “auxiliary” sequence:

$$f(\mathbf{z}_n) - f(\mathbf{x}_*) \leq \frac{2L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{(n+1)^2}.$$

For any FO method, Nesterov constructed a function f such that

$$\frac{\frac{3}{32}L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{(n+1)^2} \leq f(\mathbf{x}_n) - f(\mathbf{x}_*).$$

Thus $O(1/n^2)$ rate of FGM1 is optimal.

New results (Donghwan Kim & JF, 2016):

- Bound on convergence rate of primary sequence $\{\mathbf{x}_n\}$:

$$f(\mathbf{x}_n) - f(\mathbf{x}_*) \leq \frac{2L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{(n+2)^2}.$$

- Verifies (numerically inspired) conjecture of Drori & Teboulle (2014).

First-order (FO) method with fixed step-size coefficients:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n h_{n+1,k} \nabla f(\mathbf{x}_k)$$

- ▶ Analyze (*i.e.*, bound) convergence rate as a function of
 - ▶ number of iterations N
 - ▶ Lipschitz constant L
 - ▶ step-size coefficients $H = \{h_{n+1,k}\}$
 - ▶ initial distance to a solution: $R = \|\mathbf{x}_0 - \mathbf{x}_*\|$.
- ▶ Optimize H by minimizing the bound.
- ▶ Seek an equivalent recursive form for efficient implementation.

Ideal “universal” bound for first-order methods

For given

- number of iterations N
- Lipschitz constant L
- step-size coefficients $H = \{h_{n+1,k}\}$
- initial distance to a solution: $R = \|\mathbf{x}_0 - \mathbf{x}_*\|$,

try to bound the worst-case convergence rate of a FO method:

$$B_1(H, R, L, N, M) \triangleq \max_{f \in \mathcal{F}_L} \max_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^M} \max_{\substack{\mathbf{x}_* \in \mathcal{X}^*(f) \\ \|\mathbf{x}_0 - \mathbf{x}_*\| \leq R}} f(\mathbf{x}_N) - f(\mathbf{x}_*)$$

$$\text{such that } \mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n h_{n+1,k} \nabla f(\mathbf{x}_k), \quad n = 0, \dots, N-1.$$

Clearly for any FO method, this cost-function bound would hold:

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) \leq B_1(H, R, L, N, M).$$

For convex functions with L -Lipschitz gradients:

$$\frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{z})\|^2 \leq f(\mathbf{x}) - f(\mathbf{z}) - \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle, \quad \forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^M.$$

Drori & Teboulle (2014) use this inequality to propose a “more tractable” (finite-dimensional) relaxed bound:

$$B_2(H, R, L, N, M) \triangleq \max_{\mathbf{g}_0, \dots, \mathbf{g}_N \in \mathbb{R}^M} \max_{\delta_0, \dots, \delta_N \in \mathbb{R}} \max_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^M} \max_{\mathbf{x}_* : \|\mathbf{x}_0 - \mathbf{x}_*\| \leq R} LR\delta_N^2$$

such that
$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n h_{n+1,k} R \mathbf{g}_k, \quad n = 0, \dots, N-1,$$

$$\frac{1}{2} \|\mathbf{g}_i - \mathbf{g}_j\|^2 \leq \delta_i - \delta_j - \frac{1}{R} \langle \mathbf{g}_j, \mathbf{x}_i - \mathbf{x}_j \rangle, \quad i, j = 0, \dots, N, *$$

where $\mathbf{g}_n = \frac{1}{LR} \nabla f(\mathbf{x}_n)$ and $\delta_n = \frac{1}{LR} (f(\mathbf{x}_n) - f(\mathbf{x}_*))$.

For any FO method:

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) \leq B_1(H, R, L, N, M) \leq B_2(H, R, L, N, M)$$

However, even B_2 is as of yet unsolved.

- ▶ Drori & Teboulle (2014) further relax the bound:

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) \leq B_1(H, \dots) \leq B_2(H, \dots) \leq B_3(H, R, L, N).$$

- ▶ For given step-size coefficients H , and given number of iterations N , they use a semi-definite program (SDP) to compute B_3 numerically.
- ▶ They find numerically that for the FGM1 choice of H , the convergence bound B_3 is slightly below $\frac{2L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{(N+1)^2}$.
- ▶ This suggested that improvements on FGM1 could exist.

Drori & Teboulle (2014) also computed numerically the minimizer over H of their relaxed bound for given N using a SDP:

$$H^* = \arg \min_H B_3(H, R, L, N).$$

Numerical solution for H^* for $N = 5$ iterations: [2, Ex. 3]

$$\begin{aligned} 0. & \text{ Input: } f \in C_L^{1,1}(\mathbb{R}^d), x_0 \in \mathbb{R}^d, \\ 1. & x_1 = x_0 - \frac{1.6180}{L} f'(x_0), \\ 2. & x_2 = x_1 - \frac{0.1741}{L} f'(x_0) - \frac{2.0194}{L} f'(x_1), \\ 3. & x_3 = x_2 - \frac{0.0756}{L} f'(x_0) - \frac{0.4425}{L} f'(x_1) - \frac{2.2317}{L} f'(x_2), \\ 4. & x_4 = x_3 - \frac{0.0401}{L} f'(x_0) - \frac{0.2350}{L} f'(x_1) - \frac{0.6541}{L} f'(x_2) - \frac{2.3656}{L} f'(x_3), \\ 5. & x_5 = x_4 - \frac{0.0178}{L} f'(x_0) - \frac{0.1040}{L} f'(x_1) - \frac{0.2894}{L} f'(x_2) - \frac{0.6043}{L} f'(x_3) - \\ & \frac{2.0778}{L} f'(x_4). \end{aligned}$$

Drawbacks:

- Must choose N in advance
- Requires $O(N)$ memory for all gradient vectors $\{\nabla f(\mathbf{x}_n)\}_{n=1}^N$
- $O(N^2)$ computation for N iterations

Benefit: convergence bound (for specific N) $\approx 2 \times$ lower than for Nesterov's FGM1.

- Analytical solution for optimized step-size coefficients [8], [9]:

$$H^* : h_{n+1,k} = \begin{cases} \frac{\theta_{n-1}}{\theta_{n+1}} h_{n,k}, & k = 0, \dots, n-2 \\ \frac{\theta_{n-1}}{\theta_{n+1}} (h_{n,n-1} - 1), & k = n-1 \\ 1 + \frac{2\theta_{n-1}}{\theta_{n+1}}, & k = n. \end{cases}$$

$$\theta_n = \begin{cases} 1, & n = 0 \\ \frac{1}{2} \left(1 + \sqrt{1 + 4\theta_{n-1}^2} \right), & n = 1, \dots, N-1 \\ \frac{1}{2} \left(1 + \sqrt{1 + 8\theta_{n-1}^2} \right), & n = N. \end{cases}$$

- Analytical convergence bound for this optimized H^* :

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) \leq B_3(H^*, R, L, N) = \frac{1L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{(N+1)(N+1+\sqrt{2})}.$$

- Of course bound is $O(1/N^2)$, but constant is twice better.
- No numerical SDP needed \implies feasible for large N .
- (History: sought banded / structured lower-triangular form)

Motivation

Problem setting

Existing algorithms

- Gradient descent

- Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

- Logistic regression for machine learning

- CT image reconstruction

- Further acceleration using OS

Generalizing OGM

Summary / future work

Optimized gradient method (OGM1)

Donghwan Kim & JF (2016) also found **efficient recursive** iteration:

Initialize: $\theta_0 = 1$, $\mathbf{z}_0 = \mathbf{x}_0$

$$\mathbf{z}_{n+1} = \mathbf{x}_n - \frac{1}{L} \nabla f(\mathbf{x}_n)$$

$$\theta_n = \begin{cases} \frac{1}{2} \left(1 + \sqrt{1 + 4\theta_{n-1}^2} \right), & n = 1, \dots, N-1 \\ \frac{1}{2} \left(1 + \sqrt{1 + 8\theta_{n-1}^2} \right), & n = N \end{cases}$$

$$\mathbf{x}_{n+1} = \mathbf{z}_{n+1} + \frac{\theta_n - 1}{\theta_{n+1}} (\mathbf{z}_{n+1} - \mathbf{z}_n) + \underbrace{\frac{\theta_n}{\theta_{n+1}} (\mathbf{z}_{n+1} - \mathbf{x}_n)}_{\text{new momentum}}.$$

Reverts to Nesterov's FGM1 if the **new term** is removed.

- Very simple modification of existing Nesterov code.
- No need to solve SDP.
- Factor of 2 better bound than Nesterov's "optimal" FGM1.

(Proofs omitted.)

New version OGM1' [10], [11]

$$\mathbf{z}_{n+1} = \mathbf{x}_n - \frac{1}{L} \nabla f(\mathbf{x}_n) \quad (\text{usual GD update})$$

$$t_{n+1} = \frac{1}{2} \left(1 + \sqrt{1 + 4t_n^2} \right) \quad (\text{momentum factors})$$

$$\mathbf{x}_{n+1} = \mathbf{z}_{n+1} + \underbrace{\frac{t_n - 1}{t_{n+1}} (\mathbf{z}_{n+1} - \mathbf{x}_n)}_{\text{OGM1 momentum}}$$

- ▶ New convergence bound for *every iteration*:

$$f(\mathbf{z}_n) - f(\mathbf{x}_*) \leq \frac{1L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{(n+1)^2}.$$

- ▶ Simpler and more practical implementation.
- ▶ Need not pick N in advance.

Optimized gradient method (OGM) is optimal!

For the class of first-order (FO) methods with fixed step sizes:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{L} \sum_{k=0}^n h_{n+1,k} \nabla f(\mathbf{x}_k),$$

we optimized OGM and proved the convergence rate upper bound:

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) \leq \frac{L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{N^2}.$$

Recently Y. Drori [12] considered the class of general FO methods:

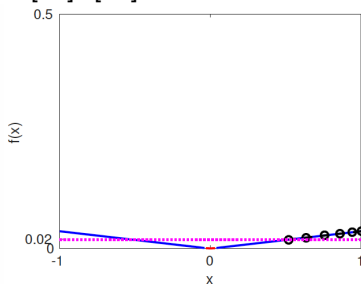
$$\mathbf{x}_{n+1} = F(\mathbf{x}_0, f(\mathbf{x}_0), \nabla f(\mathbf{x}_0), \dots, f(\mathbf{x}_n), \nabla f(\mathbf{x}_n)),$$

and showed any algorithm in this case has a function f such that

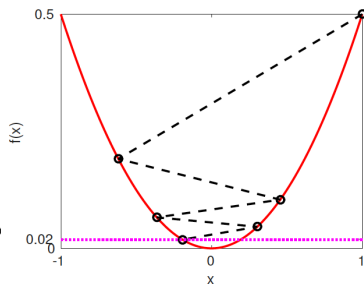
$$\frac{L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{N^2} \leq f(\mathbf{x}_N) - f(\mathbf{x}_*).$$

Thus OGM has **optimal** complexity among all FO methods!

From [10], [11], worst-case behavior is:



(c) $N = 5: f_{1,OGM}(x;5)$



(d) $N = 5: f_2(x)$

OGM has two worst-case functions (like GM):
a Huber-like function and a quadratic function.

Worst-case means:

$$f(\mathbf{x}_N) - f(\mathbf{x}_*) = \frac{LR^2}{\theta_N^2} \leq \frac{LR^2}{(N+1)(N+1+\sqrt{2})} \leq \frac{LR^2}{(N+1)^2}.$$

Motivation

Problem setting

Existing algorithms

- Gradient descent

- Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

- Logistic regression for machine learning

- CT image reconstruction

- Further acceleration using OS

Generalizing OGM

Summary / future work

To learn weights \mathbf{x} of binary classifier given feature vectors $\{\mathbf{v}_i\}$ and labels $\{y_i = \pm 1\}$:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} f(\mathbf{x}), \quad f(\mathbf{x}) = \sum_i \psi(y_i \langle \mathbf{x}, \mathbf{v}_i \rangle) + \beta \frac{1}{2} \|\mathbf{x}\|_2^2.$$

logistic:

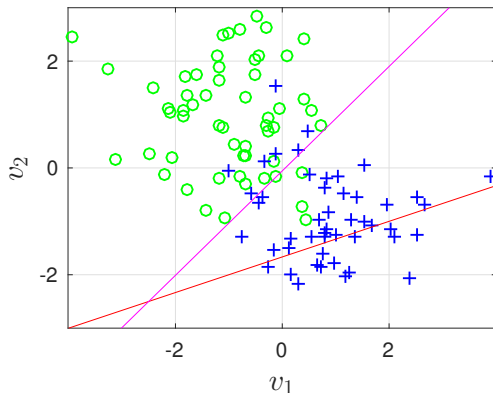
$$\psi(z) = \log(1 + e^{-z}), \quad \dot{\psi}(z) = \frac{-1}{e^z + 1}, \quad \ddot{\psi}(z) = \frac{e^z}{(e^z + 1)^2} \in \left(0, \frac{1}{4}\right].$$

Gradient $\nabla f(\mathbf{x}) = \sum_i y_i \mathbf{v}_i \dot{\psi}(y_i \langle \mathbf{x}, \mathbf{v}_i \rangle) + \beta \mathbf{x}$

Hessian is positive definite so strictly convex:

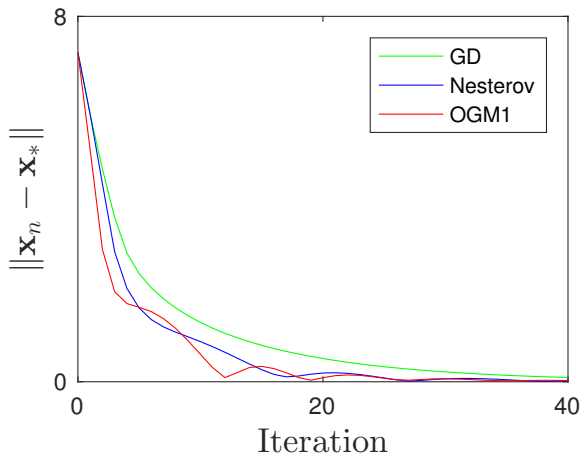
$$\nabla^2 f(\mathbf{x}) = \sum_i \mathbf{v}_i \ddot{\psi}(y_i \langle \mathbf{x}, \mathbf{v}_i \rangle) \mathbf{v}_i' + \beta \mathbf{I} \succeq \frac{1}{4} \sum_i \mathbf{v}_i \mathbf{v}_i' + \beta \mathbf{I}$$

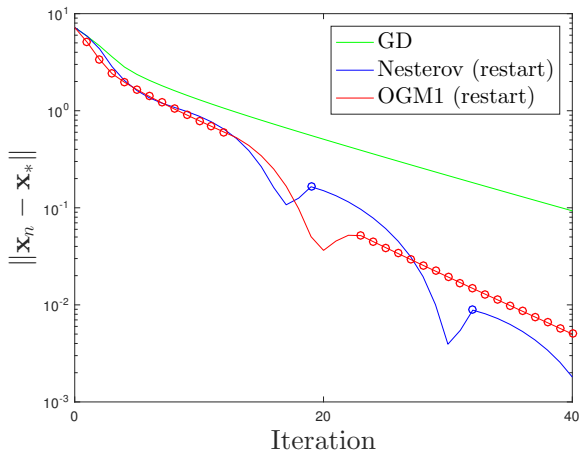
$$\implies L \triangleq \frac{1}{4} \rho \left(\sum_i \mathbf{v}_i \mathbf{v}_i' \right) + \beta \geq \max_{\mathbf{x}} \rho \left(\nabla^2 f(\mathbf{x}) \right)$$



Training data (points); initial decision boundary (red);
final decision boundary (magenta).

Numerical Results: convergence rates

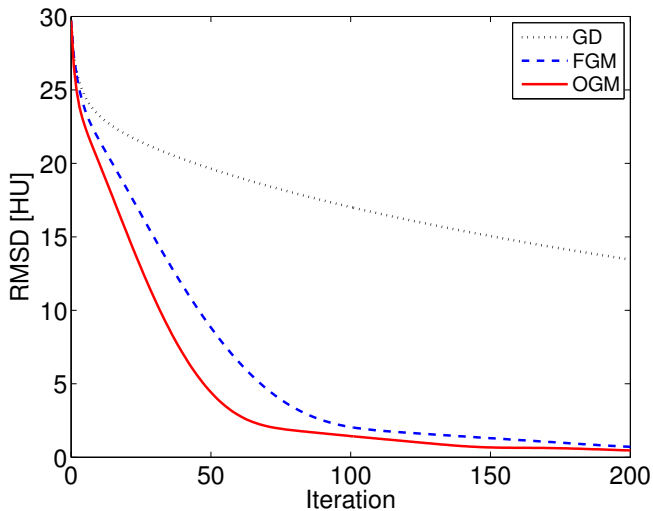




FGM restart, O'Donoghue & Candès, 2015.

How to best “restart” OGM1 is an open question.

Low-dose 2D X-ray CT image reconstruction simulation



Motivation

Problem setting

Existing algorithms

- Gradient descent

- Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

- Logistic regression for machine learning

- CT image reconstruction

- Further acceleration using OS**

Generalizing OGM

Summary / future work

- ▶ Optimization problems in image reconstruction (and machine learning) involve sums of many similar terms:

$$f(\mathbf{x}) = \sum_{m=1}^M f_m(\mathbf{x}).$$

- ▶ Approximate gradients using just one term at a time:

$$\nabla f(\mathbf{x}) \approx M \nabla f_m(\mathbf{x})$$

- ▶ Ordered subsets (OS) in tomography [15]
 - ▶ Incremental gradients in optimization / machine learning
- ▶ Combining OS with momentum dramatically accelerates!

Initialize: $\theta_0 = 1$, $\mathbf{z}_0 = \mathbf{x}_0$

(D. Kim, S. Ramani, JF, 2015) [16]

For each iteration n

For each subset $m = 1, \dots, M$

$$k = nM + m - 1$$

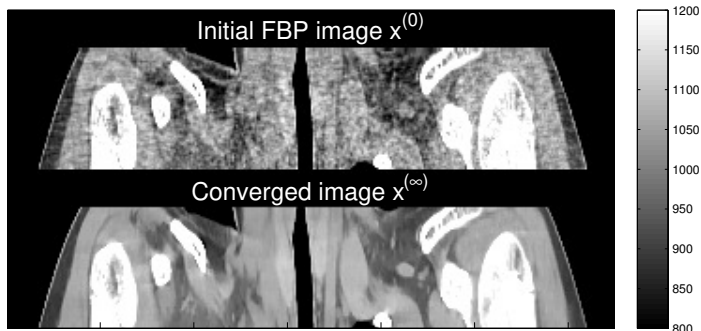
$$\mathbf{z}_{k+1} = \mathbf{x}_k - \frac{M}{L} \nabla f_m(\mathbf{x}_k) \quad (\text{usual OS update})$$

$$\theta_k = \frac{1}{2} \left(1 + \sqrt{1 + 4\theta_{k-1}^2} \right) \quad (\text{momentum factors})$$

$$\mathbf{x}_{k+1} = \mathbf{z}_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}} (\mathbf{z}_{k+1} - \mathbf{z}_k) + \underbrace{\frac{\theta_k}{\theta_{k+1}} (\mathbf{z}_{k+1} - \mathbf{x}_k)}_{\text{new momentum}}.$$

- Simple modification of existing OS code
- $\approx O(1/(Mn)^2)$ decrease of cost function f in early iterations

- 3D cone-beam helical CT scan with pitch 0.5

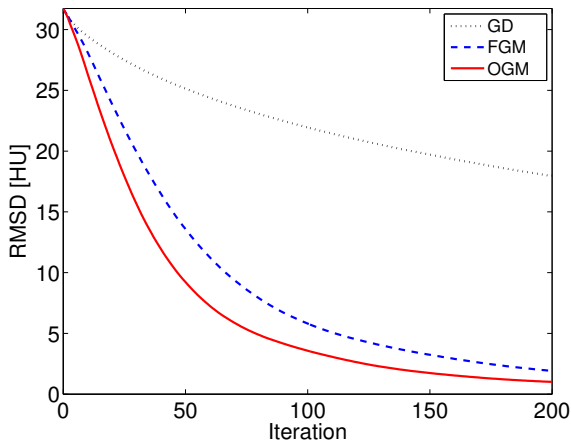


- Convergence rate in RMSD [HU], within ROI, versus iteration:

$$\text{RMSD}_{\text{ROI}}(\mathbf{x}_n) \triangleq \frac{\|x_{\text{ROI}}^{(n)} - \hat{x}_{\text{ROI}}\|_2}{\sqrt{N_{\text{ROI}}}}$$

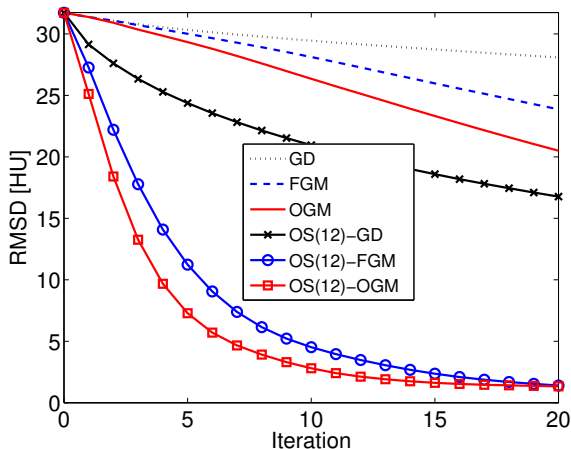
(Disclaimer: RMSD may not relate to task performance...)

Results: RMSD [HU] vs. iteration: without OS



- Computation time: **OGM** < **FGM** \ll **GD**
- **OGM** requires about $\frac{1}{\sqrt{2}}$ -times fewer iterations than **FGM** to reach the same RMSD.

Results: RMSD [HU] vs. iteration: with OS



- $M = 12$ subsets in OS algorithm.
- Proposed OS-OGM converges faster than OS-FGM.
- Computation time per iteration of all algorithms are similar.

Motivation

Problem setting

Existing algorithms

- Gradient descent

- Nesterov's "optimal" first-order method

Optimizing first-order minimization methods

Numerical examples

- Logistic regression for machine learning

- CT image reconstruction

- Further acceleration using OS

Generalizing OGM

Summary / future work

- ▶ Cost function decrease: $f(\mathbf{x}_n) - f(\mathbf{x}_*) \sim O(1/n^2)$
- ▶ Gradient norm decrease? $\|\nabla f(\mathbf{x}_n)\| \rightarrow 0$ at what rate?

Important especially for problems involving duality.

- ▶ Known bounds [17] [19]:

$$\text{GM: } \min_{0 \leq n \leq N} \|\nabla f(\mathbf{x}_n)\| = \|\nabla f(\mathbf{x}_N)\| \leq \frac{\sqrt{2}}{N} LR$$

$$\text{FGM: } \|\nabla f(\mathbf{x}_N)\| \leq \frac{2}{N} LR.$$

- ▶ New very recent bounds (DK & JF, 2016) [20], [21]:

$$\text{FGM: } \min_{0 \leq n \leq N} \|\nabla f(\mathbf{x}_n)\| \leq \frac{2\sqrt{3}}{N^{3/2}} LR$$

$$\text{OGM: } \min_{0 \leq n \leq N} \|\nabla f(\mathbf{x}_n)\| \leq \|\nabla f(\mathbf{x}_N)\| \leq \frac{\sqrt{2}}{N} LR.$$

- ▶ Can one do better than FGM?

Generalized OGM (GOGM) recursive iteration

Very recent generalization (DK & JF, 2016) [20], [21]

Input: $f \in \mathcal{F}_L$, $\mathbf{x}_0 \in \mathbb{R}^N$, $\mathbf{z}_0 = \mathbf{x}_0$, $t_0 \in (0, 1]$.
for $n = 0, 1, \dots$

$$\mathbf{z}_{n+1} = \mathbf{x}_n - \frac{1}{L} \nabla f(\mathbf{x}_n)$$

$$t_{n+1} > 0 \text{ s.t. } t_{n+1}^2 \leq T_{n+1} \triangleq \sum_{k=0}^{n+1} t_k \quad (\text{momentum factors})$$

$$\begin{aligned} \mathbf{x}_{n+1} = \mathbf{z}_{n+1} &+ \frac{(T_n - t_n)t_{n+1}}{T_{n+1}t_n} (\mathbf{z}_{n+1} - \mathbf{z}_n) \\ &+ \frac{(2t_n^2 - T_n)t_{n+1}}{T_{n+1}t_n} (\mathbf{z}_{n+1} - \mathbf{x}_n). \end{aligned}$$

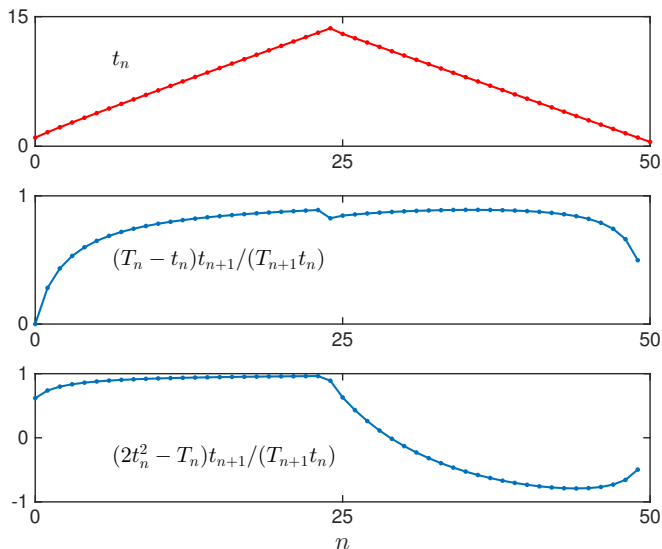
- ▶ Simple implementation
- ▶ Best choice of factors t_n (in terms of gradient norm decrease)?

Optimized choice of momentum factors (for decreasing gradient norm) (DK & JF, 2016) [20], [21] :

$$t_n \triangleq \begin{cases} 1, & n = 0, \\ \frac{1}{2} \left(1 + \sqrt{1 + 4t_{n-1}^2} \right), & n = 0, \dots, \lfloor N/2 \rfloor - 1, \\ (N - n + 1)/2, & n = \lfloor N/2 \rfloor, \dots, N. \end{cases}$$

Dubbed “OGM-OG” for OGM with optimized gradients.

Optimized parameters for OGM-OG



- ▶ Convergence bound for cost function for OGM-OG:

$$f(\mathbf{z}_N) - f(\mathbf{x}_*) \leq \frac{2L \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{N^2}.$$

- ▶ Same as Nesterov's FGM.
- ▶ Convergence bound for gradient norm is best known:

$$\min_{0 \leq n \leq N} \|\nabla f(\mathbf{z}_n)\| \leq \min_{0 \leq n \leq N} \|\nabla f(\mathbf{x}_n)\| \leq \frac{\sqrt{6}}{N^{3/2}} LR.$$

- ▶ $\sqrt{2}$ better than FGM's *smallest* gradient norm bound.
- ▶ Variations that do not require choosing N in advance, but that have slightly larger constants in bounds.
- ▶ Derivation uses relaxations that are not tight.
- ▶ Is $N^{3/2}$ best possible? What is best possible constant?

Summary of (fast?) gradient decreasing FO methods

From [20], [21]:

Algorithm	Asymptotic convergence rate bound		Require selecting N in advance
	Cost function	Gradient norm	
GM	$\frac{1}{4}N^{-1}$	$\sqrt{2}N^{-1}$	No
FGM	$2N^{-2}$	$2\sqrt{3}N^{-\frac{3}{2}}$	No
OGM	N^{-2}	$\sqrt{2}N^{-1}$	No
OGM-H	$4N^{-2}$	$4N^{-\frac{3}{2}}$	Yes
OGM-OG	$2N^{-2}$	$\sqrt{6}N^{-\frac{3}{2}}$	Yes
OGM- a ($a > 2$)	$\frac{a}{2}N^{-2}$	$\frac{a\sqrt{6}}{2\sqrt{a-2}}N^{-\frac{3}{2}}$	No
OGM- $a=4$	$2N^{-2}$	$2\sqrt{3}N^{-\frac{3}{2}}$	

Numerical examples are work-in-progress.

Composite cost function:

$$\arg \min_{\mathbf{x}} F(\mathbf{x}), \quad F(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x})$$

$f(\mathbf{x})$: convex, smooth with Lipschitz gradient

$g(\mathbf{x})$: convex but possibly (usually) non-smooth

Examples:

- $g(\mathbf{x}) = \|\mathbf{x}\|_1$
- $g(\mathbf{x})$ characteristic function of a convex constraint

Fast iterative soft thresholding algorithm (FISTA) (Beck & Teboulle, 2009) [22]

AKA “fast proximal gradient method” (FPGM)

Simple recursive iteration with $O(1/n^2)$ cost function convergence rate

DK & JF, 2016 [23], [24]

Algorithm	Asymptotic convergence rate bound		Require selecting N in advance
	Cost function ($\times LR^2$)	Proximal gradient ($\times LR$)	
PGM	$\frac{1}{2}N^{-1}$	$2N^{-1}$	No
FPGM [5]	$2N^{-2}$	$2N^{-1}$	No
FPGM- σ ($0 < \sigma < 1$) [22]	$\frac{2}{\sigma^2}N^{-2}$	$\frac{2\sqrt{3}}{\sigma^2} \sqrt{\frac{1+\sigma}{1-\sigma}} N^{-\frac{3}{2}}$	No
FPGM- $\sigma=0.78$	$3.3N^{-2}$	$16.2N^{-\frac{3}{2}}$	
FPGM-H	$8N^{-2}$	$5.7N^{-\frac{3}{2}}$	Yes
FPGM-OPG	$4N^{-2}$	$4.9N^{-\frac{3}{2}}$	Yes
FPGM- a ($a > 2$)	aN^{-2}	$\frac{a\sqrt{6}}{\sqrt{a-2}} N^{-\frac{3}{2}}$	No
FPGM- $a=4$	$4N^{-2}$	$6.9N^{-\frac{3}{2}}$	

FPGM with “optimized proximal gradient” (FPGM-OPG).

Best known bound on proximal gradient convergence rate.

- ▶ New optimized first-order minimization algorithm (optimal!)
- ▶ Simple implementation akin to Nesterov's FGM
- ▶ Analytical converge rate bound
- ▶ Bound on cost function decrease is $2\times$ better than Nesterov

Future work

- Constraints
- Non-smooth cost functions, e.g., ℓ_1
- Tighter bounds
- Strongly convex case
- Asymptotic / local convergence rates
- Incremental gradients
- Stochastic gradient descent
- Adaptive restart
- Distributed computation
- Low-dose 3D X-ray CT image reconstruction

- [1] R. C. Fair, "On the robust estimation of econometric models," *Ann. Econ. Social Measurement*, vol. 2, 667–77, Oct. 1974.
- [2] Y. Drori and M. Teboulle, "Performance of first-order methods for smooth convex minimization: A novel approach," *Mathematical Programming*, vol. 145, no. 1-2, 451–82, Jun. 2014.
- [3] A. B. Taylor, J. M. Hendrickx, and François. Glineur, "Smooth strongly convex interpolation and exact worst-case performance of first- order methods," *Mathematical Programming*, 2016.
- [4] B. T. Polyak, *Introduction to optimization*. New York: Optimization Software Inc, 1987.
- [5] J. Barzilai and J. Borwein, "Two-point step size gradient methods," *IMA J. Numerical Analysis*, vol. 8, no. 1, 141–8, 1988.
- [6] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$," *Dokl. Akad. Nauk. USSR*, vol. 269, no. 3, 543–7, 1983.
- [7] —, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, 127–52, May 2005.
- [8] D. Kim and J. A. Fessler, *Optimized first-order methods for smooth convex minimization*, arxiv 1406.5468, 2014.
- [9] —, "Optimized first-order methods for smooth convex minimization," *Mathematical Programming*, vol. 159, no. 1, 81–107, Sep. 2016.
- [10] —, *On the convergence analysis of the optimized gradient methods*, arxiv 1510.08573, 2015.
- [11] —, "On the convergence analysis of the optimized gradient methods," *J. Optim. Theory Appl.*, 2016, Submitted.
- [12] Y. Drori, *The exact information-based complexity of smooth convex minimization*, arxiv 1606.01424, 2016.

- [13] D. Böhning and B. G. Lindsay, "Monotonicity of quadratic approximation algorithms," *Ann. Inst. Stat. Math.*, vol. 40, no. 4, 641–63, Dec. 1988.
- [14] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Found. Comp. Math.*, vol. 15, no. 3, 715–32, Jun. 2015.
- [15] H. Erdoğan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Phys. Med. Biol.*, vol. 44, no. 11, 2835–51, Nov. 1999.
- [16] D. Kim, S. Ramani, and J. A. Fessler, "Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, 167–78, Jan. 2015.
- [17] Y. Nesterov, *How to make the gradients small*, *Optima* 88, 2012.
- [18] A. Beck and M. Teboulle, "A fast dual proximal gradient algorithm for convex minimization and applications," *Operations Research Letters*, vol. 42, no. 1, 1–6, Jan. 2014.
- [19] I. Necoara and A. Patrascu, "Iteration complexity analysis of dual first order methods for conic convex programming," *Optimization Methods and Software*, vol. 31, no. 3, 645–78, 2016.
- [20] D. Kim and J. A. Fessler, *Generalizing the optimized gradient method for smooth convex minimization*, arxiv 1607.06764, 2016.
- [21] —, "Generalizing the optimized gradient method for smooth convex minimization," *Mathematical Programming*, 2016, Submitted.
- [22] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Im. Proc.*, vol. 18, no. 11, 2419–34, Nov. 2009.
- [23] D. Kim and J. A. Fessler, *Another look at the "Fast iterative shrinkage/Thresholding algorithm (FISTA)*, arxiv 1608.03861, 2016.
- [24] —, "Another look at the "Fast iterative shrinkage/Thresholding algorithm (FISTA)",," *SIAM J. Optim.*, 2016, Submitted.