

First-Order Methods

Stephen J. Wright¹

²Computer Sciences Department,
University of Wisconsin-Madison.

PCMI, July 2016

Thanks

Mario Figueiredo.

Smooth Convex Functions

Consider $\min_{x \in \mathbb{R}^n} f(x)$, with f smooth and convex.

Usually assume $mI \preceq \nabla^2 f(x) \preceq LI$, $\forall x$, with $0 \leq m \leq L$.

Thus L is a Lipschitz constant of ∇f :

$$\|\nabla f(x) - \nabla f(z)\| \leq L\|x - z\|,$$

and

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2.$$

If $m > 0$, then f is m -strongly convex and

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|_2^2.$$

Define **conditioning** (or condition number) as $\kappa := L/m$.

What's the Setup?

We consider **iterative algorithms**: generate $\{x_k\}$, $k = 0, 1, 2, \dots$ from

$$x_{k+1} = \Phi(x_k) \quad \text{or} \quad x_{k+1} = \Phi(x_k, x_{k-1}) \quad \text{or} \quad x_{k+1} = \Phi(x_k, x_{k-1}, \dots, x_1, x_0).$$

For now, assume we can evaluate $f(x_t)$ and $\nabla f(x_t)$ at each iteration. Some of the techniques we discuss are extendible to more general situations:

- nonsmooth f ;
- f not available (or too expensive to evaluate exactly);
- only an *estimate* of the gradient is available;
- a constraint $x \in \Omega$, usually for a simple Ω (e.g. ball, box, simplex);
- nonsmooth regularization; *i.e.*, instead of simply $f(x)$, we want to minimize $f(x) + \tau\psi(x)$.

We focus on algorithms that can be adapted to those scenarios.

Steepest Descent

Minimizer x^* of f is characterized by $\nabla f(x^*) = 0$.

At a point for which $\nabla f(x) \neq 0$, can get decrease in f by moving in any direction d such that $d^T \nabla f(x) < 0$. Proof is from Taylor's theorem:

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^T d + O(\alpha^2) < f(x), \quad \text{for } \alpha \text{ sufficiently small.}$$

Among all d with $\|d\| = 1$, the minimizer of $d^T \nabla f(x)$ is attained at $d = -\nabla f(x)$. This is the **steepest descent** direction.

Even when f is not convex, the direction d with $d^T \nabla f(x) = 0$ will decrease f from any point for which $\nabla f(x) \neq 0$. Algorithms that take “reasonable” steps along $d = -\nabla f(x)$ at each iteration cannot accumulate at points $\bar{x}$ for which $\nabla f(\bar{x}) \neq 0$ — can always escape from a neighborhood of such points.

Steepest Descent

Steepest descent (a.k.a. gradient descent):

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \text{for some } \alpha_k > 0.$$

Different ways to select an appropriate α_k .

- 1 Interpolating scheme with safeguarding to identify an approximate minimizing α_k .
- 2 Backtrack. Try $\bar{\alpha}$, $\frac{1}{2}\bar{\alpha}$, $\frac{1}{4}\bar{\alpha}$, $\frac{1}{8}\bar{\alpha}$, ... until sufficient decrease in f .
- 3 Don't test for function decrease; use rules based on L and m .
- 4 Set α_k based on experience with similar problems. Or adaptively.

Analysis for 1 and 2 usually yields global convergence at unspecified rate. The “greedy” strategy of getting good decrease in the current search direction may lead to better practical results.

Analysis for 3: Focuses on convergence rate, and leads to accelerated multi-step methods.

Fixed Steps

By elementary use of Taylor's theorem, and since $\nabla^2 f(x) \preceq LI$,

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \|\nabla f(x_k)\|_2^2.$$

For $\alpha_k \equiv 1/L$,
$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2,$$

thus
$$\|\nabla f(x_k)\|^2 \leq 2L[f(x_k) - f(x_{k+1})]$$

Summing over first $T - 1$ iterates ($k = 0, 1, \dots, T - 1$) and telescoping the sum,

$$\sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq 2L[f(x_0) - f(x_T)].$$

It follows that $\nabla f(x_k) \rightarrow 0$ if f is bounded below.

Convergence Rates

From the sum above we have that

$$T \min_{k=0,1,\dots,T-1} \|\nabla f(x_k)\|^2 \leq \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq 2L[f(x_0) - f(x_T)],$$

and so

$$\min_{k=0,1,\dots,T-1} \|\nabla f(x_k)\| \leq \sqrt{\frac{2L[f(x_0) - f(x_T)]}{T}}.$$

“Smallest gradient encountered in first T iterations shrinks like $1/\sqrt{T}$.”
This result doesn't require convexity!

For convergence of function values $\{f(x_k)\}$ to their optimal value f^* in the **convex case**, we have the following remarkably bound:

$$f(x_T) - f^* \leq \frac{L}{2T} \|x_0 - x^*\|_2^2.$$

Proof on following slides!

Proof of $1/T$ Convergence of $\{f(x_T)\}$

For any solution x^* , have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 \\ &\leq f^* + \nabla f(x_k)^T (x_k - x^*) - \frac{1}{2L} \|\nabla f(x_k)\|^2 \quad (\text{convexity}) \\ &= f(x^*) + \frac{L}{2} \left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right) \\ &= f(x^*) + \frac{L}{2} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2). \end{aligned}$$

By summing over $k = 0, 1, 2, \dots, T-1$, we have

$$\begin{aligned} \sum_{k=0}^{T-1} (f(x_{k+1}) - f^*) &\leq \frac{L}{2} \sum_{k=0}^{T-1} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) \\ &= \frac{L}{2} (\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) \\ &\leq \frac{L}{2} \|x^0 - x^*\|^2. \end{aligned}$$

Since $\{f(x^k)\}$ is nonincreasing, have

$$f(x_T) - f(x^*) \leq \frac{1}{T} \sum_{k=0}^{T-1} (f(x_{k+1}) - f_*) \leq \frac{L}{2T} \|x_0 - x^*\|_2^2$$

as required. That's it!

Strongly convex: Linear Rate

From strong convexity condition, we have for any z :

$$f(z) \geq f(x_k) + \nabla f(x_k)^T (z - x_k) + \frac{m}{2} \|z - x_k\|^2.$$

By minimizing both sides w.r.t. z we obtain

$$f(x^*) \geq f(x_k) - \frac{1}{2m} \|\nabla f(x_k)\|^2,$$

so that

$$\|\nabla f(x_k)\|^2 \geq 2m(f(x_k) - f(x^*)). \quad (1)$$

Recall too that for step $\alpha_k \equiv 1/L$ we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2.$$

Subtract $f(x^*)$ from both sides of this expression and use (1):

$$(f(x_{k+1}) - f(x^*)) \leq \left(1 - \frac{m}{L}\right) (f(x_k) - f(x^*)).$$

A **linear (geometric)** rate!

A Word on Convergence Rates

Typical rates of convergence to zero for sequences such as $\{\|\nabla f(x_k)\|\}$, $\{f(x^k) - f^*\}$, and $\{\|x^k - x^*\|\}$ are

$$\phi_k \leq \frac{C_1}{\sqrt{k}}, \frac{C_2}{k}, \frac{C_3}{k^2} \quad (\text{sublinear})$$

$$\phi_{k+1} \leq (1 - c)\phi_k \text{ for some } c \in (0, 1) \quad (\text{linear})$$

$$\phi_{k+1} = o(\phi_k) \quad (\text{superlinear}).$$

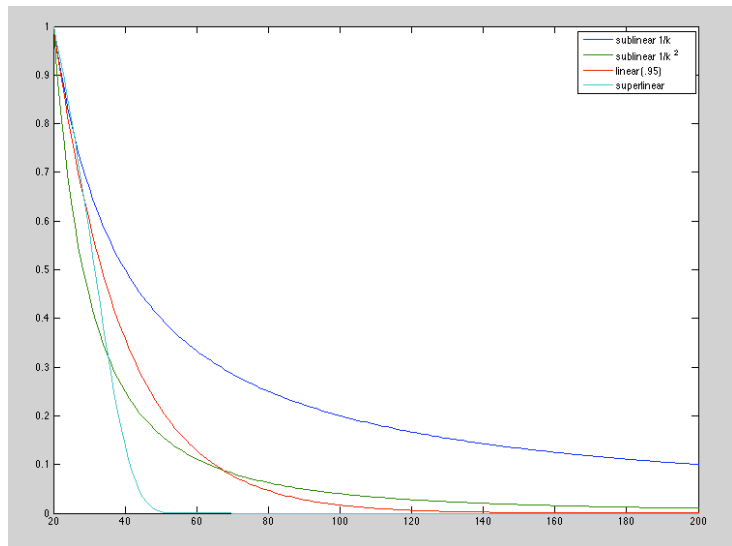
To achieve $\phi_T \leq \epsilon$ for some small positive tolerance ϵ , need

$$T = O(1/\epsilon^2), \quad T = O(1/\epsilon), \quad T = O(1/\sqrt{\epsilon}) \quad \text{for sublinear rates,}$$

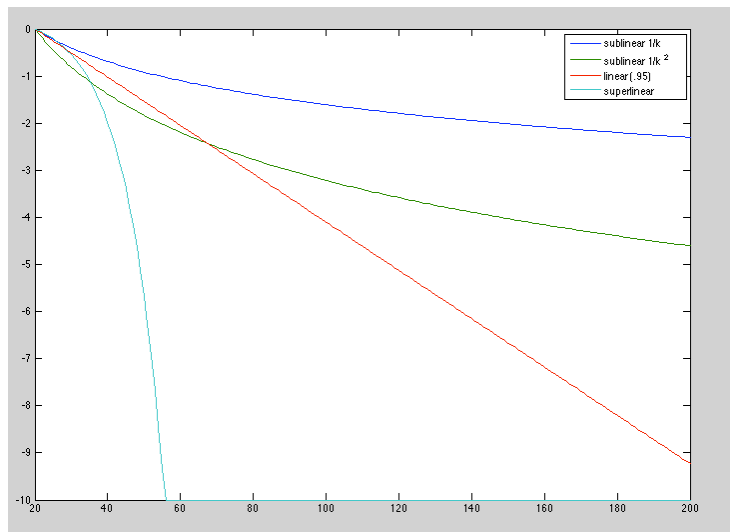
$$T = O\left(\frac{1}{c} \log \epsilon\right), \quad \text{for linear rate.}$$

Question: For a quadratic convergence rate $\phi_{k+1} \leq C\phi_k^2$, how many iterations are required to obtain $\phi_T \leq \epsilon$?

Convergence Rates: Standard Plots



Convergence Rates: Log Plots



Linear convergence without strong convexity

The linear convergence analysis depended on two bounds:

$$f(x_{k+1}) \leq f(x_k) - a_1 \|\nabla f(x_k)\|^2, \quad (2)$$

$$\|\nabla f(x_k)\|^2 \geq a_2(f(x_k) - f(x^*)), \quad (3)$$

for some positive a_1, a_2 . In fact, many algorithms that use first derivatives, or crude estimates of first derivatives (as in stochastic gradient or coordinate descent) satisfy a bound like (2).

We derived (3) from strong convexity, but it also holds for interesting cases that are **not strongly convex**.

(3) is a special case of a Kurdyka-Lojasewicz (KL) property, which holds in many interesting situations — even for nonconvex f , near a local min.

The KL property holds when f grows quadratically from its solution set:

$$f(x) - f^* \geq a_3 \text{dist}(x, \text{solution set})^2, \quad \text{for some } a_3 > 0.$$

Allows nonunique solution. Proof:

$$\begin{aligned} f(x) - f^* &\leq -\nabla f(x)^T (x - x^*) \\ &\leq \|\nabla f(x)\| \|x - x^*\| \\ &\leq \|\nabla f(x)\| \sqrt{(f(x) - f^*)/a_3}. \end{aligned}$$

So obtain by rearrangement that

$$\|\nabla f(x)\|^2 \leq a_3(f(x) - f^*).$$

KL also holds when $f(x) = \sum_{i=1}^m h(a_i^T x)$, where $h : \mathbb{R} \rightarrow \mathbb{R}$ is strongly convex, **even when $m < n$** , in which case $\nabla^2 f(x)$ is singular. This form of f arises in **Empirical Risk Minimization (ERM)**.

The $1/k^2$ Speed Limit

Nesterov (2004) gives a simple example of a smooth function for which no method that generates iterates of the form $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ can converge at a rate faster than $1/k^2$, at least for its first $n/2$ iterations.

Note that $x_{k+1} \in x_0 + \text{span}(\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_k))$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ & & \ddots & \ddots & \ddots & & \\ 0 & \dots & & & 0 & -1 & 2 \end{bmatrix}, \quad e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and set $f(x) = (1/2)x^T A x - e_1^T x$. The solution has $x^*(i) = 1 - i/(n+1)$.

If we start at $x_0 = 0$, each $\nabla f(x_k)$ has nonzeros only in its first k entries. Hence, $x_{k+1}(i) = 0$ for $i = k+1, k+2, \dots, n$. Can show

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}.$$

Descent Directions and Line Search

Consider iteration scheme

$$x_{k+1} = x_k + \alpha_k d_k, \quad k = 0, 1, 2, \dots,$$

where d_k makes an acute angle with $-\nabla f(x_k)$, that is,

$$-d_k^T \nabla f(x_k) \geq \bar{\epsilon} \|\nabla f(x_k)\| \|d_k\|. \quad (4)$$

We impose **weak Wolfe conditions** on steplength α_k :

$$f(x_k + \alpha d_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T d_k, \quad (5a)$$

$$\nabla f(x_k + \alpha d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k. \quad (5b)$$

where $0 < c_1 < c_2 < 1$. (Typically $c_1 = .001$, $c_2 = .5$.)

- (5a) is a **sufficient decrease condition**;
- (5b) ensures that the step is not too short.

Second weak Wolfe condition

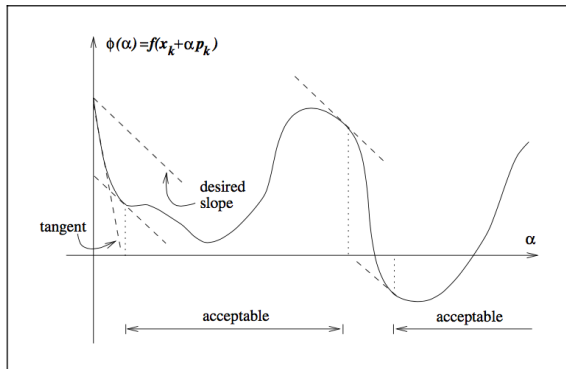


Figure 3.4 The curvature condition.

Convergence under Weak Wolfe

From condition (5b) and the Lipschitz property for ∇f , we have

$$-(1 - c_2)\nabla f(x_k)^T d_k \leq [\nabla f(x_k + \alpha_k d_k) - \nabla f(x_k)]^T d_k \leq L\alpha_k \|d_k\|^2,$$

and thus

$$\alpha_k \geq -\frac{(1 - c_2)}{L} \frac{\nabla f(x_k)^T d_k}{\|d_k\|^2}.$$

Substituting into (5a), and using (4), we have

$$\begin{aligned} f(x_{k+1}) &= f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k \\ &\leq f(x_k) - \frac{c_1(1 - c_2)}{L} \frac{(\nabla f(x_k)^T d_k)^2}{\|d_k\|^2} \\ &\leq f(x_k) - \frac{c_1(1 - c_2)}{L} \bar{\epsilon}^2 \|\nabla f(x_k)\|^2. \end{aligned}$$

Thus the decrease in f per iteration is a multiple of $\|\nabla f(x_k)\|^2$, just as in vanilla steepest descent with fixed steps. We thus get the same sublinear and linear convergence results.

Backtracking

Try $\alpha_k = \bar{\alpha}, \frac{\bar{\alpha}}{2}, \frac{\bar{\alpha}}{4}, \frac{\bar{\alpha}}{8}, \dots$ until the **sufficient decrease** condition is satisfied.

No need to check the second Wolfe condition: the α_k thus identified is “within striking distance” of an α that’s too large — so it is not too short.

Backtracking is widely used in applications, but **doesn't work on nonsmooth problems**, or when f is not available / too expensive.

Can show again that the decrease in f at each iteration is a multiple of $\|\nabla f(x^k)\|^2$, so the usual rates apply.

Exact minimizing α_k : Faster rate?

Question: does taking α_k as the exact minimizer of f along $-\nabla f(x_k)$ yield better rate of linear convergence?

Consider $f(x) = \frac{1}{2}x^T A x$ (thus $x^* = 0$ and $f(x^*) = 0$.)

We have $\nabla f(x_k) = A x_k$. Exactly minimizing w.r.t. α_k ,

$$\alpha_k = \arg \min_{\alpha} \frac{1}{2}(x_k - \alpha A x_k)^T A (x_k - \alpha A x_k) = \frac{x_k^T A^2 x_k}{x_k^T A^3 x_k} \in \left[\frac{1}{L}, \frac{1}{m} \right]$$

Thus

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2} \frac{(x_k^T A^2 x_k)^2}{(x_k^T A x_k)(x_k^T A^3 x_k)},$$

so, defining $z_k := A x_k$, we have

$$\frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} \leq 1 - \frac{\|z_k\|^4}{(z_k^T A^{-1} z_k)(z_k^T A z_k)}.$$

Exact minimizing α_k : Faster rate?

Using Kantorovich inequality:

$$(z^T A z)(z^T A^{-1} z) \leq \frac{(L + m)^2}{4Lm} \|z\|^4.$$

Thus

$$\frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} \leq 1 - \frac{4Lm}{(L + m)^2} \approx 1 - \frac{4m}{L},$$

Only a small factor of improvement in the linear rate over constant steplength.

Convergence of Iterates x_k

Can we say something about the rate of convergence of $\{x_k\}$ to x^* ? That is, convergence of $\|x_k - x^*\|$ or $\text{dist}(x_k, \text{minimizing set})$ to zero?

In the weakly convex case, not much! $f(x^k) - f^*$ can be small while x^k is still far from x^* .

If strong convexity holds, we have

$$f(x_k) \geq f(x^*) + \frac{1}{2}m\|x^k - x^*\|^2,$$

so that

$$\|x_k - x^*\| \leq \sqrt{\frac{2}{m}(f(x_k) - f^*)}.$$

So we can derive convergence rates on $\|x_k - x^*\|$ from those of $f(x_k) - f^*$.

Can get rates under weaker conditions e.g. quadratic growth

$$f(x) - f^* \geq a_3 \text{dist}(x, \text{solution set})^2, \quad \text{for some } a_3 > 0.$$

The slow linear rate is typical!

Not just a pessimistic bound! In the strongly convex case, complexity to achieve $f(x_T) - f^* \leq \epsilon(f(x_0) - f^*)$ is $O((L/m) \log \epsilon)$.



Accelerated First-Order Methods

Can we get faster rates (e.g. faster linear rates for strongly convex, faster sublinear rates for general convex) while still using only first-order information?

YES! The key idea is **MOMENTUM**. Search direction depends on the latest gradient $-\nabla f(x_k)$ and also on the **search direction at iteration $k-1$** , which encodes gradient information from all earlier iterations.

Several popular methods use momentum:

- Heavy-ball method
- Nesterov's accelerated gradient
- Conjugate gradient (linear and nonlinear).

Heavy Ball:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Nesterov's optimal method:

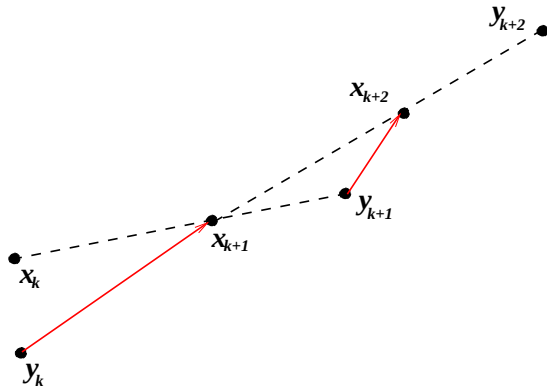
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k + \beta_k(x_k - x_{k-1})) + \beta_k(x_k - x_{k-1}).$$

Typically $\alpha_k \approx 1/L$ and $\beta_k \approx 1$.

Can rewrite Nesterov by introducing an intermediate sequence $\{y_k\}$:

$$\begin{aligned} y_k &= x_k + \beta_k(x_k - x_{k-1}), \\ x_{k+1} &= x_k - \alpha_k \nabla f(y_k) + \beta_k(x_k - x_{k-1}). \end{aligned}$$

Nesterov, illustrated



Separates the “gradient descent” and “momentum” step components.

Accelerated Gradient Convergence

Typical convergence:

Weakly convex $m = 0$: $f(x_k) - f^* = O(1/k^2)$;

Strongly convex $m > 0$: $f(x_k) - f^* \leq M \left(1 - c\sqrt{\frac{m}{L}}\right)^k [f(x_0) - f^*]$,

for some modest positive c .

- Approach can be extended to regularized functions $f(x) + \lambda\psi(x)$: Beck and Teboulle (2009b).
- Partial-gradient approaches (stochastic gradient, coordinate descent) can be accelerated in similar ways.

Heavy Ball

Consider heavy-ball applied to a convex quadratic:

$$f(x) = \frac{1}{2}x^T Qx,$$

where Q is symmetric positive definite with eigenvalues

$$0 < m = \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 = L.$$

The minimizer is clearly $x^* = 0$.

Heavy ball applied to this function is

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) = x_k - \alpha Qx_k + \beta(x_k - x_{k-1}).$$

Analyze by defining a composite iterate vector:

$$w_k := \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \end{bmatrix} = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$

Thus

$$w_k = Tw_{k-1}, \quad T := \begin{bmatrix} (1 + \beta)I - \alpha Q & -\beta I \\ I & 0 \end{bmatrix}.$$

Multistep Methods: The Heavy-Ball

Matrix T has same eigenvalues as

$$\begin{bmatrix} -\alpha\Lambda + (1 + \beta)I & -\beta I \\ I & 0 \end{bmatrix}, \quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

Can rearrange this matrix to get 2×2 blocks on the diagonal:

$$T_i := \begin{bmatrix} (1 + \beta) - \alpha\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}.$$

Get eigenvalues by solving quadratics:

$$u^2 - (1 + \beta - \alpha\lambda_i)u + \beta = 0,$$

Eigenvalues are all complex provided that $(1 + \beta - \alpha\lambda_i)^2 - 4\beta < 0$, which happens when

$$\beta \in \left((1 - \sqrt{\alpha\lambda_i})^2, (1 + \sqrt{\alpha\lambda_i})^2 \right).$$

Heavy Ball, continued

Thus the eigenvalues of T are all complex:

$$\begin{aligned}\bar{\lambda}_{i,1} &= \frac{1}{2} \left[(1 + \beta - \alpha\lambda_i) + i\sqrt{4\beta - (1 + \beta - \alpha\lambda_i)^2} \right], \\ \bar{\lambda}_{i,2} &= \frac{1}{2} \left[(1 + \beta - \alpha\lambda_i) - i\sqrt{4\beta - (1 + \beta - \alpha\lambda_i)^2} \right].\end{aligned}$$

All eigenvalues have magnitude β !

Thus can do an eigenvalue decomposition $T = VSV^{-1}$, where S is diagonal with entries $\bar{\lambda}_{i,1}, \bar{\lambda}_{i,2}, i = 1, 2, \dots, n$.

The recurrence becomes

$$w_k = Tw_{k-1} = T^k w_0 = V S^k V^{-1} w_0.$$

Thus we have

$$\|V^{-1}w_k\| = \|S^k V^{-1}w_0\| \leq \|S^k\| \|V^{-1}w_0\| = \beta^k \|V^{-1}w_0\|.$$

Note that this does not imply monotonic decrease in $\|w_k\|$, only in the scaled norm $\|V^{-1}w_k\|$.

Heavy-Ball: Optimal choice of α and β

We want to minimize β , but need β to satisfy

$$\beta \in \left((1 - \sqrt{\alpha \lambda_i})^2, (1 + \sqrt{\alpha \lambda_i})^2 \right), \quad \text{with } \lambda_i \in [m, L],$$

which is satisfied when

$$\beta = \min(|1 - \sqrt{\alpha m}|, |1 - \sqrt{\alpha L}|)^2$$

Choose α to make the two quantities on the right-hand side identical:

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{m})^2} \Rightarrow 1 - \sqrt{\alpha m} = -(1 - \sqrt{\alpha L}) = \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}}.$$

It follows that

$$\beta = \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}} = 1 - \frac{2}{\sqrt{L/m} + 1}.$$

Caution!

The heavy ball analysis is elementary and powerful.

- The asymptotic rate is better than for Nesterov.
- The rate is as good as the classical conjugate gradient method for $Ax = b$. (In fact, the analysis techniques are very similar.)

But we need to note a few things!

- It depends on knowledge of m and L in order to make the right choices of α and β .
- It doesn't extend neatly from quadratic to **nonlinear** f .
- We can't prove contraction for the weakly convex case $m = 0$.

Exercise: Repeat this analysis for Nesterov's optimal method (again for convex quadratic f).

Summary: Linear Convergence, Strictly Convex f

Defining $\kappa = L/m$, rates are approximately:

- Steepest descent: Linear rate approx $\left(1 - \frac{2}{\kappa}\right)$;
- Heavy-ball: Linear rate approx $\left(1 - \frac{2}{\sqrt{\kappa}}\right)$.

Big difference! To reduce $\|x_k - x^*\|$ by a factor ϵ , need k large enough that

$$\left(1 - \frac{2}{\kappa}\right)^k \leq \epsilon \iff k \geq \frac{\kappa}{2} |\log \epsilon| \quad (\text{steepest descent})$$

$$\left(1 - \frac{2}{\sqrt{\kappa}}\right)^k \leq \epsilon \iff k \geq \frac{\sqrt{\kappa}}{2} |\log \epsilon| \quad (\text{heavy-ball})$$

A factor of $\sqrt{\kappa}$ difference; e.g. if $\kappa = 1000$, need ~ 30 times fewer steps.

Conjugate Gradient

Basic **conjugate gradient** (CG) step is

$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k = -\nabla f(x_k) + \gamma_k p_{k-1}.$$

Can be identified with heavy-ball, with $\beta_k = \frac{\alpha_k \gamma_k}{\alpha_{k-1}}$.

However, CG can be implemented in a way that doesn't require knowledge (or estimation) of L and m .

- Choose α_k to (approximately) minimize f along p_k ;
- Choose γ_k by a variety of formulae (Fletcher-Reeves, Polak-Ribiere, etc), all of which are equivalent if f is convex quadratic. e.g.

$$\gamma_k = \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2}$$

Conjugate Gradient

Nonlinear CG: Variants include Fletcher-Reeves, Polak-Ribiere, Hestenes.

Restarting periodically with $p_k = -\nabla f(x_k)$ is useful (e.g. every n iterations, or when p_k is not a descent direction).

For **quadratic** f , convergence analysis is based on eigenvalues of A and Chebyshev polynomials, min-max arguments. Get

- **Finite termination** in as many iterations as there are distinct eigenvalues;
- **Asymptotic linear convergence** with rate approx $1 - \frac{2}{\sqrt{\kappa}}$.
(like heavy-ball.)

(Nocedal and Wright, 2006, Chapter 5)

Nesterov (1983) describes a method that requires L and m and makes adaptive choices of α_k, β_k .

Initialize: Choose $x_0, \alpha_0 \in (0, 1)$; set $y_0 \leftarrow x_0$.

Iterate: $x_{k+1} \leftarrow y_k - \frac{1}{L} \nabla f(y_k)$; (*short-step*)

find $\alpha_{k+1} \in (0, 1)$: $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + \frac{\alpha_{k+1}}{\kappa}$;

set $\beta_k = \frac{\alpha_k(1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$;

set $y_{k+1} \leftarrow x_{k+1} + \beta_k(x_{k+1} - x_k)$.

Still works for weakly convex ($m = 0$). Just set $\kappa = \infty$ in the scheme above.

Convergence Results: Nesterov

If $\alpha_0 \geq 1/\sqrt{\kappa}$, have

$$f(x_k) - f(x^*) \leq c_1 \min \left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k, \frac{4L}{(\sqrt{L} + c_2 k)^2} \right),$$

where constants c_1 and c_2 depend on x_0 , α_0 , L .

- Linear convergence “heavy-ball” rate for strongly convex f ;
- $1/k^2$ sublinear rate otherwise.

In the special case of $\alpha_0 = 1/\sqrt{\kappa}$, this scheme yields

$$\alpha_k \equiv \frac{1}{\sqrt{\kappa}}, \quad \beta_k \equiv 1 - \frac{2}{\sqrt{\kappa} + 1}.$$

Beck and Teboulle (2009a) propose a similar algorithm, with a fairly short and elementary analysis (though still not intuitive).

Initialize: Choose x_0 ; set $y_1 = x_0$, $t_1 = 1$;

Iterate: $x_k \leftarrow y_k - \frac{1}{L} \nabla f(y_k)$;

$$t_{k+1} \leftarrow \frac{1}{2} \left(1 + \sqrt{1 + 4t_k^2} \right);$$

$$y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}} (x_k - x_{k-1}).$$

For (weakly) convex f , converges with $f(x_k) - f(x^*) \sim 1/k^2$.

When L is not known, increase an estimate of L until it's big enough.

Beck and Teboulle (2009a) do the convergence analysis in 2-3 pages; elementary, but “technical.”

A Non-Monotone Gradient Method: Barzilai-Borwein

Barzilai and Borwein (1988) (BB) proposed an unusual choice of α_k .
Allows f to increase (sometimes a lot) on some steps: **non-monotone**.

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad \alpha_k := \arg \min_{\alpha} \|s_k - \alpha z_k\|^2,$$

where

$$s_k := x_k - x_{k-1}, \quad z_k := \nabla f(x_k) - \nabla f(x_{k-1}).$$

Explicitly, we have

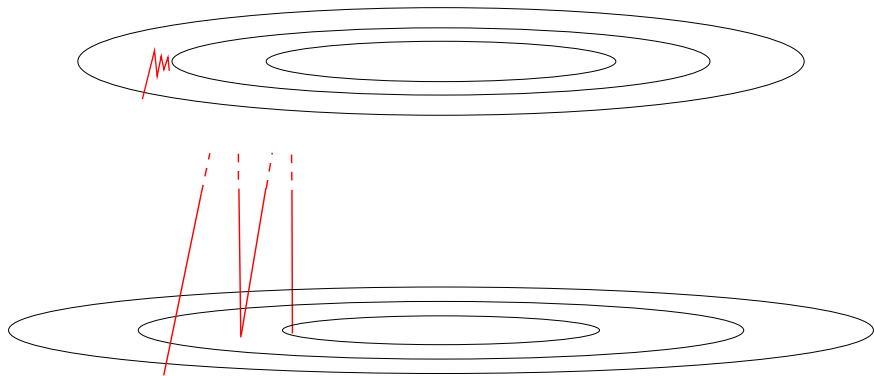
$$\alpha_k = \frac{s_k^T z_k}{z_k^T z_k}.$$

Note that for $f(x) = \frac{1}{2}x^T A x$, we have

$$\alpha_k = \frac{s_k^T A s_k}{s_k^T A^2 s_k} \in \left[\frac{1}{L}, \frac{1}{m} \right].$$

BB can be viewed as a quasi-Newton method, with the Hessian approximated by $\alpha_k^{-1} I$.

Comparison: BB vs Greedy Steepest Descent



There Are Many BB Variants

- use $\alpha_k = s_k^T s_k / s_k^T z_k$ in place of $\alpha_k = s_k^T z_k / z_k^T z_k$;
- alternate between these two formulae;
- hold α_k constant for a number (2, 3, 5) of successive steps;
- take α_k to be the steepest descent step from the [previous](#) iteration.

Nonmonotonicity appears essential to performance. Some variants get global convergence by requiring a sufficient decrease in f over the worst of the last M (say 10) iterates.

The original 1988 analysis in BB's paper is nonstandard and illuminating (just for a 2-variable quadratic).

In fact, most analyses of BB and related methods are nonstandard, and consider only special cases. The precursor of such analyses is Akaike (1959). More recently, see Ascher, Dai, Fletcher, Hager and others.

Extending to the Constrained Case: $x \in \Omega$

How to change these methods to handle the **constraint** $x \in \Omega$?
(assuming that Ω is a **closed convex set**)

Some algorithms and theory stay much the same,

...if we can involve the constraint $x \in \Omega$ explicitly in the subproblems.

Example: Nesterov's constant step scheme requires just one calculation to be changed from the unconstrained version.

Initialize: Choose $x_0, \alpha_0 \in (0, 1)$; set $y_0 \leftarrow x_0$.

Iterate: $x_{k+1} \leftarrow \arg \min_{y \in \Omega} \frac{1}{2} \|y - [y_k - \frac{1}{L} \nabla f(y_k)]\|_2^2$;
find $\alpha_{k+1} \in (0, 1)$: $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + \frac{\alpha_{k+1}}{\kappa}$;
set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$;
set $y_{k+1} \leftarrow x_{k+1} + \beta_k(x_{k+1} - x_k)$.

Convergence theory is unchanged.

Regularized Optimization

How to change these methods to handle **regularized optimization**?

$$\min_x f(x) + \lambda\psi(x),$$

where f is convex and smooth, while ψ is convex but usually **nonsmooth**.

Often, all that is needed is to change the update step to

$$x_{k+1} = \arg \min_x \frac{1}{2} \|x - \Phi(x_k)\|_2^2 + \alpha_k \lambda \psi(x). \quad (6)$$

where $\Phi(x_k)$ could be a steepest descent step

$$\Phi(x_k) = x_k - \alpha_k \nabla f(x_k),$$

or something more complicated (such as heavy ball, or some other accelerated method). When $\lambda = 0$, we have simply $x_{k+1} = \Phi(x_k)$, so this reverts to the standard first-order methods described above.

(6) is the **shrinkage/thresholding** step; how to solve it with a nonsmooth ψ ? That's the topic of the following slides.

Another Motivation

We can view shrinking / thresholding alternatively as a first-order subproblem with a quadratic prox term.

$$\begin{aligned}x_{k+1} &= \arg \min_x \frac{1}{2} \|x - (x_k - \alpha_k \nabla f(x_k))\|_2^2 + \alpha_k \lambda \psi(x) \\&= \arg \min_x -\nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 + \lambda \psi(x),\end{aligned}$$

where we divided by α_k in the second expression and dropped the term that's independent of x .

This subproblem:

- makes a linear approximation to f at x_k ;
- incorporates a quadratic prox term with weight $1/\alpha_k$;
- incorporates the regularization term $\lambda \psi(x)$ **explicitly, without modification**.

This approach makes sense when the **subproblem is easy to solve**. This is true in a number of interesting cases.

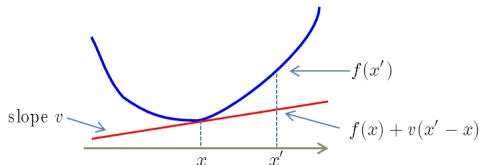
Reminder: Subgradients

Subgradients generalize gradients for general convex functions:

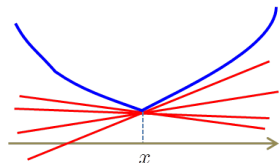
$$v \text{ is a subgradient of } f \text{ at } x \text{ if } f(x') \geq f(x) + v^T(x' - x)$$

Subdifferential: $\partial f(x) = \{\text{all subgradients of } f \text{ at } x\}$

If f is differentiable, $\partial f(x) = \{\nabla f(x)\}$



linear lower bound



nondifferentiable case

Subgradients satisfy a **monotonicity property**: If $a \in \partial f(x)$ and $b \in \partial f(y)$, then $(a - b)^T(x - y) \geq 0$.

A Key Tool: Moreau's Proximity Operators

Moreau (1962) proximity operator

$$\hat{x} \in \arg \min_x \frac{1}{2} \|x - y\|_2^2 + \psi(x) =: \text{prox}_\psi(y)$$

...well defined for convex ψ , since $\|\cdot - y\|_2^2$ is coercive and strictly convex.

Example: $\text{prox}_{\tau|\cdot|}(y) = \text{soft}(y, \tau) = \text{sign}(y) \max\{|y| - \tau, 0\}$

Block separability: $x = (x_1, \dots, x_N)$ (a partition of the components of x)

$$\psi(x) = \sum_i \psi_i(x_i) \Rightarrow (\text{prox}_\psi(y))_i = \text{prox}_{\psi_i}(y_i)$$

Relationship with subdifferential: $z = \text{prox}_\psi(y) \Leftrightarrow z - y \in \partial\psi(z)$

Resolvent: $z = \text{prox}_\psi(y) \Leftrightarrow 0 \in \partial\psi(z) + (z - y) \Leftrightarrow y \in (\partial\psi + I)z$

$$\text{prox}_\psi(y) = (\partial\psi + I)^{-1}y$$

Prox operators and the Moreau envelope

Moreau envelope:

$$M_{\lambda,\psi}(y) := \frac{1}{\lambda} \inf_x \left\{ \frac{1}{2} \|x - y\|_2^2 + \lambda \psi(x) \right\}$$

The minimizer in $M_{\lambda,\psi}(y)$ is achieved at $\text{prox}_{\lambda\psi}(y)$.

By optimality properties, we have

$$y - \text{prox}_{\lambda\psi}(y) \in \lambda \partial\psi(\text{prox}_{\lambda\psi}(y)).$$

$M_{\lambda,\psi}(y)$ can be viewed as a smoothing of ψ , differentiable everywhere:

$$\nabla M_{\lambda,\psi}(y) = \frac{1}{\lambda} (y - \text{prox}_{\lambda\psi}(y)).$$

By monotonicity of ∂ , together with optimality condition above, can show that prox is a contraction, that is,

$$\|\text{prox}_{\lambda\psi}(y) - \text{prox}_{\lambda\psi}(z)\| \leq \|y - z\|.$$

Important Proximity Operators

- **Soft-thresholding** is the proximity operator of the ℓ_1 norm.
- Consider the **indicator** ι_S of a **convex set** S ;

$$\text{prox}_{\iota_S}(u) = \arg \min_x \frac{1}{2} \|x - u\|_2^2 + \iota_S(x) = \arg \min_{x \in S} \frac{1}{2} \|x - u\|_2^2 = P_S(u)$$

...the **Euclidean projection** on S .

- Squared Euclidean norm (separable, smooth): **Exercise!**
- Euclidean norm (not separable, nonsmooth):

$$\text{prox}_{\tau \|\cdot\|_2}(y) = \begin{cases} \frac{y}{\|y\|_2} (\|y\|_2 - \tau), & \text{if } \|y\|_2 > \tau \\ 0 & \text{if } \|y\|_2 \leq \tau \end{cases}$$

More Proximity Operators

$\phi(x)$	$\text{prox}_{\phi}x$
i $\ell_{[\underline{\omega}, \overline{\omega}]}(x)$	$P_{[\underline{\omega}, \overline{\omega}]}x$
ii $\sigma_{[\underline{\omega}, \overline{\omega}]}(x) = \begin{cases} \underline{\omega}x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \overline{\omega}x & \text{otherwise} \end{cases}$	$\text{soft}_{[\underline{\omega}, \overline{\omega}]}(x) = \begin{cases} x - \underline{\omega} & \text{if } x < \underline{\omega} \\ 0 & \text{if } x \in [\underline{\omega}, \overline{\omega}] \\ x - \overline{\omega} & \text{if } x > \overline{\omega} \end{cases}$
iii $\begin{matrix} \psi(x) + \sigma_{[\underline{\omega}, \overline{\omega}]}(x) \\ \psi \in \Gamma_0(\mathbb{R}) \text{ differentiable at } 0 \\ \psi'(0) = 0 \end{matrix}$	$\text{prox}_{\psi}(\text{soft}_{[\underline{\omega}, \overline{\omega}]}(x))$
iv $\max\{ x - \omega, 0\}$	$\begin{cases} x & \text{if } x < \omega \\ \text{sign}(x)\omega & \text{if } \omega \leq x \leq 2\omega \\ \text{sign}(x)(x - \omega) & \text{if } x > 2\omega \end{cases}$
v $\kappa x ^q$	$\text{sign}(x)p$, where $p \geq 0$ and $p + q\kappa p^{q-1} = x $
vi $\begin{cases} \kappa x^2 & \text{if } x \leq \omega/\sqrt{2\kappa} \\ \omega\sqrt{2\kappa} x - \omega^2/2 & \text{otherwise} \end{cases}$	$\begin{cases} x/(2\kappa + 1) & \text{if } x \leq \omega(2\kappa + 1)/\sqrt{2\kappa} \\ x - \omega\sqrt{2\kappa}\text{sign}(x) & \text{otherwise} \end{cases}$
vii $\omega x + \tau x ^2 + \kappa x ^q$	$\text{sign}(x)\text{prox}_{\kappa \cdot ^q/(2\tau+1)}\left(\frac{\max\{ x - \omega, 0\}}{2\tau + 1}\right)$
viii $\omega x - \ln(1 + \omega x)$	$(2\omega)^{-1} \text{sign}(x) \left(\omega x - \omega^2 - 1 + \sqrt{[\omega x - \omega^2 - 1]^2 + 4\omega x } \right)$
Many others!	
ix $\begin{cases} \omega x^{-q} & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} x - \omega & \text{if } x > \omega \\ 0 & \text{if } x \in [0, \omega] \\ +\infty & \text{otherwise} \end{cases}$
x $\begin{cases} \omega x^{-q} & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} p > 0 \\ \text{such that } p^{q+2} - xp^{q+1} = \omega q \end{cases}$
xi $\begin{cases} \omega x^{-q} & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} p > 0 \\ \text{such that } p^{q+2} - xp^{q+1} = \omega q \end{cases}$
xii $\begin{cases} x \ln(x) & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ +\infty & \text{otherwise} \end{cases}$	$W(e^{x-1})$, where W is the Lambert W-function
xiii $\begin{cases} -\ln(x - \underline{\omega}) + \ln(-\underline{\omega}) & \text{if } x \in [\underline{\omega}, 0] \\ -\ln(\overline{\omega} - x) + \ln(\overline{\omega}) & \text{if } x \in [0, \overline{\omega}] \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} \frac{1}{2}(x + \underline{\omega} + \sqrt{ x - \underline{\omega} ^2 + 4}) & \text{if } x < 1/\underline{\omega} \\ \frac{1}{2}(x + \overline{\omega} - \sqrt{ x - \overline{\omega} ^2 + 4}) & \text{if } x > 1/\overline{\omega} \\ 0 & \text{otherwise} \end{cases}$ (see Figure 1)
xiv $\begin{cases} -\kappa \ln(x) + \tau x^2/2 + \alpha x & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$	$\frac{1}{2(1+\tau)}(x - \alpha + \sqrt{ x - \alpha ^2 + 4\kappa(1+\tau)})$
xv $\begin{cases} -\kappa \ln(x) + \alpha x + \omega x^{-1} & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} p > 0 \\ \text{such that } p^3 + (\alpha - x)p^2 - \kappa p = \omega \end{cases}$
xvi $\begin{cases} -\kappa \ln(x) + \omega x^q & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} p > 0 \\ \text{such that } q\omega p^q + p^2 - xp = \kappa \end{cases}$
xvii $\begin{cases} -\underline{\kappa} \ln(x - \underline{\omega}) - \overline{\kappa} \ln(\overline{\omega} - x) & \text{if } x \in [\underline{\omega}, \overline{\omega}] \\ +\infty & \text{otherwise} \end{cases}$	$\begin{cases} p \in [\underline{\omega}, \overline{\omega}] \\ \text{such that } p^3 - (\underline{\omega} + \overline{\omega} + x)p^2 + (\underline{\omega}\overline{\omega} - \underline{\kappa} - \overline{\kappa} + (\underline{\omega} + \overline{\omega})x)p = \underline{\omega}\overline{\omega}x - \underline{\omega}\overline{\kappa} - \overline{\omega}\underline{\kappa} \end{cases}$

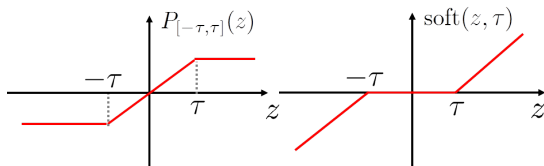
(Combettes and Pesquet, 2011)

From Conjugates to Proximity Operators

Notice that $|u| = \sup_{x \in [-1,1]} x^T u = \sigma_{[-1,1]}(u)$, thus $|\cdot|^* = \iota_{[-1,1]}$.

Using Moreau's decomposition, we easily derive the soft-threshold:

$$\text{prox}_{\tau|\cdot|} = 1 - \text{prox}_{\iota_{[-\tau,\tau]}} = 1 - P_{[-\tau,\tau]} = \text{soft}(\cdot, \tau)$$



Conjugate of a norm: if $f(x) = \tau \|x\|_p$ then $f^* = \iota_{\{x: \|x\|_q \leq \tau\}}$,

where $\frac{1}{q} + \frac{1}{p} = 1$ (a **Hölder pair**, or **Hölder conjugates**).

That is, $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms:

$$\|z\|_q = \sup\{x^T z : \|x\|_p \leq 1\} = \sup_{x \in B_p(1)} x^T z = \sigma_{B_p(1)}(z)$$

From Conjugates to Proximity Operators

- Proximity of norm:

$$\text{prox}_{\tau\|\cdot\|_p} = I - P_{B_q(\tau)}$$

where $B_q(\tau) = \{x : \|x\|_q \leq \tau\}$ and $\frac{1}{q} + \frac{1}{p} = 1$.

- Example:** computing $\text{prox}_{\|\cdot\|_\infty}$ (notice ℓ_∞ is not separable):

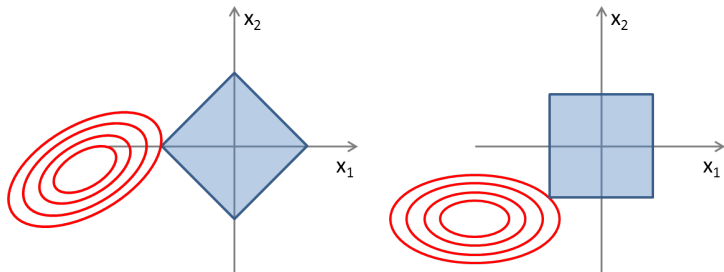
Since $\frac{1}{\infty} + \frac{1}{1} = 1$,

$$\text{prox}_{\tau\|\cdot\|_\infty} = I - P_{B_1(\tau)}$$

... the proximity operator of ℓ_∞ norm is the residual of the projection on an ℓ_1 ball.

- Projection on ℓ_1 ball has **no closed form**, but there are **efficient (linear cost) algorithms** (Brucker, 1984), (Maculan and de Paula, 1989).

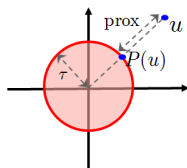
Whereas ℓ_1 promotes sparsity, ℓ_∞ promotes equality (in absolute value).



From Conjugates to Proximity Operators

The dual of the ℓ_2 norm is the ℓ_2 norm.

$$\text{prox}_{\tau\|\cdot\|_2}(u) = u - P_{\{x:\|x\|_2\leq\tau\}}(u)$$



$$= u - \begin{cases} u & \Leftarrow \|u\|_2 \leq \tau \\ \tau u / \|u\|_2 & \Leftarrow \|u\|_2 > \tau \end{cases}$$

$$= \frac{u}{\|u\|_2} \max\{0, \|u\|_2 - \tau\}$$

vector soft thresholding

Matrix Nuclear Norm and its Prox Operator

- Recall the trace/nuclear norm: $\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i$.
- The dual of a Schatten p -norm is a Schatten q -norm, with $\frac{1}{q} + \frac{1}{p} = 1$. Thus, the dual of the nuclear norm is the spectral norm:

$$\|X\|_\infty = \max \{ \sigma_1, \dots, \sigma_{\min\{m,n\}} \}.$$

- If $Y = U\Lambda V^T$ is the SVD of Y , we have

$$\begin{aligned} \text{prox}_{\tau\|\cdot\|_*}(Y) &= U\Lambda V^T - P_{\{X: \max\{\sigma_1, \dots, \sigma_{\min\{m,n\}}\} \leq \tau\}}(U\Lambda V^T) \\ &= U \text{soft}(\Lambda, \tau) V^T. \end{aligned}$$

Atomic Norms: A Unified View

vectors

matrices

norm	prox	atomic set	norm	prox	atomic set
ℓ_1 $\ x\ _1$	<u>component soft thresholding</u>	$\mathcal{A} = \{\pm e_i\}$ $ \mathcal{A} = 2N$	nuclear $\ X\ _*$	singular value thresholding	$\mathcal{A} =$ set of all rank 1, norm 1 matrices
ℓ_∞ $\ x\ _\infty$	<u>residual of projection on ℓ_1 ball</u>	$\mathcal{A} = \{\pm 1\}^N$ $ \mathcal{A} = 2^N$	spectral $\ X\ _2$	<u>residual of s.v. proj. on ℓ_1 ball</u>	$\mathcal{A} =$ set of all orthogonal matrices
ℓ_2 $\ x\ _2$	<u>vector soft thresholding</u>	$\mathcal{A} =$ set of all vectors with norm 1 $ \mathcal{A} = \infty$	Frobenius $\ X\ _F$	<u>matrix soft threshold.</u>	$\mathcal{A} =$ all matrices of unit Frobenius norm.

Another Use of Fenchel-Legendre Conjugates

- The original problem: $\min_x f(x) + \psi(x)$
- Often this has the form: $\min_x g(Ax) + \psi(x)$
- Using the definition of conjugate $g(Ax) = \sup_u u^T Ax - g^*(u)$

$$\begin{aligned}\min_x g(Ax) + \psi(x) &= \inf_x \sup_u u^T Ax - g^*(u) + \psi(x) \\ &= \sup_u (-g^*(u)) + \inf_x u^T Ax + \psi(x) \\ &= \sup_u (-g^*(u)) - \underbrace{\sup_x -x^T A^T u - \psi(x)}_{\psi^*(-A^T u)} \\ &= -\inf_u g^*(u) + \psi^*(-A^T u)\end{aligned}$$

- The dual $\inf_u g^*(u) + \psi^*(-A^T u)$ is sometimes easier to handle.

Basic Proximal-Gradient Algorithm

Use basic structure:

$$x_k = \arg \min_x \|x - \Phi(x_k)\|_2^2 + \psi(x).$$

with $\Phi(x_k)$ a simple gradient descent step, thus

$$x_{k+1} = \text{prox}_{\alpha_k \psi}(x_k - \alpha_k \nabla f(x_k))$$

This approach goes by many names, such as

- “proximal gradient algorithm” (PGA),
- “iterative shrinkage/thresholding” (IST),
- “forward-backward splitting” (FBS)

It has been reinvented several times in different communities: optimization, partial differential equations, convex analysis, signal processing, machine learning.

Convergence of Prox-Gradient

$$x_{k+1} = \text{prox}_{\alpha_k \psi}(x_k - \alpha_k \nabla f(x_k)).$$

Proof makes use of “gradient map” defined by

$$G_\alpha(x) := \frac{1}{\alpha} (x - \text{prox}_{\alpha \psi}(x - \alpha \nabla f(x))). \quad (7)$$

Can rewrite the step taken at iteration k :

$$x^{k+1} = x^k - \alpha_k G_{\alpha_k}(x^k) \Leftrightarrow G_{\alpha_k} = \frac{1}{\alpha_k} (x^k - x^{k+1}). \quad (8)$$

Lemma

Suppose that ψ is closed convex function, ∇f has Lipschitz constant L .

- (a) $G_\alpha(x) \in \nabla f(x) + \partial \psi(x - \alpha G_\alpha(x))$.
- (b) *For any z , and any $\alpha \in (0, 1/L]$, we have that*

$$\phi(x - \alpha G_\alpha(x)) \leq \phi(z) + G_\alpha(x)^T (x - z) - \frac{\alpha}{2} \|G_\alpha(x)\|^2.$$

Proof of (a)

From optimality property of the prox-operator, which is

$$y - \text{prox}_{\lambda\psi}(y) \in \lambda\partial\psi(\text{prox}_{\lambda\psi}(y)),$$

we have

$$(x - \alpha\nabla f(x)) - \text{prox}_{\alpha\psi}(x - \alpha\nabla f(x)) \in \alpha\partial\psi(\text{prox}_{\alpha\psi}(x - \alpha\nabla f(x))).$$

Now substitute from $\text{prox}_{\alpha\psi}(x - \alpha\nabla f(x)) = x - \alpha G_\alpha(x)$, to obtain

$$0 \in \alpha\partial\psi(x - \alpha G_\alpha(x)) - \alpha(G_\alpha(x) - \nabla f(x)),$$

from which (a) follows when we divide by α .

Proof of (b)

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2.$$

Set $y = x - \alpha G_\alpha(x)$, for any $\alpha \in (0, 1/L]$, to get

$$\begin{aligned} f(x - \alpha G_\alpha(x)) &\leq f(x) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{L\alpha^2}{2} \|G_\alpha(x)\|^2 \\ &\leq f(x) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{\alpha}{2} \|G_\alpha(x)\|^2. \end{aligned} \quad (9)$$

(Second inequality uses $\alpha \in (0, 1/L]$.) By convexity of f and ψ , for any z and any $v \in \partial\psi(x - \alpha G_\alpha(x))$ we have

$$f(z) \geq f(x) + \nabla f(x)^T(z - x) \quad (10a)$$

$$\psi(z) \geq \psi(x - \alpha G_\alpha(x)) + v^T(z - (x - \alpha G_\alpha(x))). \quad (10b)$$

From (a) we have $v = (G_\alpha(x) - \nabla f(x)) \in \partial\psi(x - \alpha G_\alpha(x))$, so by substituting in (10) and also using (9) we have the following...

Proof of (b), continued

$$\begin{aligned}\phi(x - \alpha G_\alpha(x)) &= f(x - \alpha G_\alpha(x)) + \psi(x - \alpha G_\alpha(x)) \\ &\leq f(x) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{\alpha}{2} \|G_\alpha(x)\|^2 + \psi(x - \alpha G_\alpha(x)) \quad (\text{from (9)}) \\ &\leq f(z) + \nabla f(x)^T (x - z) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{\alpha}{2} \|G_\alpha(x)\|^2 \\ &\quad + \psi(z) + (G_\alpha(x) - \nabla f(x))^T (x - \alpha G_\alpha(x) - z) \quad (\text{from (10)}) \\ &= f(z) + \psi(z) + G_\alpha(x)^T (x - z) - \frac{\alpha}{2} \|G_\alpha(x)\|^2,\end{aligned}$$

for any $\alpha \in (0, 1/L]$, where the last equality follows from cancellation of several terms in the previous line.

Sublinear Convergence

Denote $\phi(x) = f(x) + \psi(x)$ with minimizer x^* (not necessarily unique).

Main convergence result:

Theorem

If $\alpha_k \equiv 1/L$, have

$$\phi(x^T) - \phi^* \leq \frac{L\|x^0 - x^*\|^2}{2T}, \quad T = 1, 2, \dots$$

Use Lemma 1 (b) to show decrease of $\{\phi(x^k)\}$ and $\|x^k - x^*\|$. Set $x = z = x^k$ and $\alpha = \alpha_k$ and use (8) to obtain

$$\phi(x^{k+1}) = \phi(x^k - \alpha_k G_{\alpha_k}(x^k)) \leq \phi(x^k) - \frac{\alpha_k}{2} \|G_{\alpha_k}(x^k)\|^2,$$

showing decrease in ϕ .

For decrease in $\|x - x^*\|$, set $x = x^k$, $\alpha = \alpha_k$, and $z = x^*$ in Lemma 1:

$$\begin{aligned} 0 &\leq \phi(x^{k+1}) - \phi^* \\ &= \phi(x^k - \alpha_k G_{\alpha_k}(x^k)) - \phi^* \\ &\leq G_{\alpha_k}^T(x^k - x^*) - \frac{\alpha_k}{2} \|G_{\alpha_k}(x^k)\|^2 \\ &= \frac{1}{2\alpha_k} \left(\|x^k - x^*\|^2 - \|x^k - x^* - \alpha_k G_{\alpha_k}(x^k)\|^2 \right) \\ &= \frac{1}{2\alpha_k} \left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right), \end{aligned} \tag{11}$$

from which $\|x^{k+1} - x^*\| \leq \|x^k - x^*\|$ follows.

Set $\alpha_k = 1/L$ in (11), and sum over $k = 0, 1, 2, \dots, T-1$, obtain

$$\sum_{k=0}^{T-1} (\phi(x^{k+1}) - \phi^*) \leq \frac{L}{2} \left(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2 \right) \leq \frac{L}{2} \|x^0 - x^*\|^2.$$

By monotonicity of $\{\phi(x^k)\}$, we have

$$T(\phi(x^T) - \phi^*) \leq \sum_{k=0}^{T-1} (\phi(x^{k+1}) - \phi^*).$$

Result follows by combining these last two expressions.

Proximal-Gradient Algorithm: Quadratic Case

- Consider the **quadratic** case (of great interest): $f(x) = \frac{1}{2}\|Bx - b\|_2^2$.
- Here, $\nabla f(x) = B^T(Bx - b)$ and the IST/PGA/FBS algorithm is

$$x_{k+1} = \text{prox}_{\alpha_k \psi}(x_k - \alpha_k B^T(Bx - b))$$

can be implemented with only matrix-vector multiplications with B and B^T .

This is a **very important** feature in large-scale applications, such as image processing, where **fast algorithms** exist for computing these products (e.g. fast Fourier transforms or wavelet transforms), but these **matrices cannot be formed and stored** explicitly.

- In this case, some more refined convergence results are available.
- Even more refined results are available if $\psi(x) = \|x\|_1$

More on IST/FBS/PGA for the ℓ_2 - ℓ_1 Case

- Problem: $\hat{x} \in G = \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|Bx - b\|_2^2 + \tau \|x\|_1$ (recall $B^T B \preceq LI$)
- IST/FBS/PGA becomes $x_{k+1} = \text{soft}(x_k - \alpha B^T (Bx_k - b), \alpha\tau)$
with $\alpha < 2/L$.
- The zero set: $\mathcal{Z} \subseteq \{1, \dots, n\} : \hat{x} \in G \Rightarrow \hat{x}_{\mathcal{Z}} = 0$
- Zeros are found in a finite number of iterations (Hale et al., 2008):
after a finite number of iterations, we have $(x_k)_{\mathcal{Z}} = 0$.
- After that, if $B_{\mathcal{Z}}^T B_{\mathcal{Z}} \succeq ml$, with $m > 0$ (thus $\kappa(B_{\mathcal{Z}}^T B_{\mathcal{Z}}) = L/m$):

$$\|x_{k+1} - \hat{x}\|_2 \leq \frac{1 - \kappa}{1 + \kappa} \|x_k - \hat{x}\|_2 \quad (\text{linear convergence})$$

for the optimal choice $\alpha = 2/(L + m)$. (Weaker condition suffices for linear convergence of $\{f(x_k)\}$; see above.)

FISTA with prox operations

- Recall that FISTA — *fast iterative shrinkage-thresholding algorithm* — ((Beck and Teboulle, 2009a), based on (Nesterov, 1983)) is a heavy-ball-type **acceleration** of IST:

Initialize: Choose $\alpha \leq 1/L$, x_0 ; set $y_1 = x_0$, $t_1 = 1$;

Iterate: $x_k \leftarrow \text{prox}_{\tau\alpha\psi}(y_k - \alpha\nabla f(y_k))$;
 $t_{k+1} \leftarrow \frac{1}{2} \left(1 + \sqrt{1 + 4t_k^2} \right)$;
 $y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$.

- Acceleration:**

$$\text{FISTA: } f(x_k) - f(\hat{x}) \sim O\left(\frac{1}{k^2}\right) \quad \text{IST: } f(x_k) - f(\hat{x}) \sim O\left(\frac{1}{k}\right).$$

- When L is not known, increase an estimate of L until it's big enough.

Heavy Ball Acceleration: TwIST

- TwIST (*two-step iterative shrinkage-thresholding* (Bioucas-Dias and Figueiredo, 2007)) is a **heavy-ball-type acceleration** of IST, for

$$\min_x \frac{1}{2} \|Bx - b\|_2^2 + \tau\psi(x)$$

- Iterations (with $\alpha < 2/L$)

$$x_{k+1} = (\gamma - \beta)x_k + (1 - \gamma)x_{k-1} + \beta \operatorname{prox}_{\alpha\tau\psi}(x_k - \alpha B^T(Bx - b))$$

- Analysis in the strongly convex case: $mI \preceq B^T B \preceq LI$, with $m > 0$. Conditioning (as above) $\kappa = L/m < \infty$.
- Optimal parameters: $\gamma = \rho^2 + 1$, $\beta = \frac{2\alpha}{m+L}$, where $\rho = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$, yield linear convergence

$$\|x_{k+1} - \hat{x}\|_2 \leq \frac{1 - \sqrt{\kappa}}{1 + \sqrt{\kappa}} \|x_k - \hat{x}\|_2 \quad \left(\text{versus } \frac{1-\kappa}{1+\kappa} \text{ for IST} \right)$$

Acceleration via Larger Steps: SpaRSA

- The standard step-size $\alpha_k \leq 2/L$ in IST is **too timid**
- The **SpARSA** (**s**parse **r**econstruction **b**y **s**eparable **a**pproximation) framework proposes **bolder choices of α_k** (Wright et al., 2009):
 - ✓ Barzilai-Borwein (see above), to mimic Newton steps — or at least get the scaling right.
 - ✓ keep increasing α_k until monotonicity is violated: backtrack.
- Convergence to critical points (minima in the convex case) is guaranteed for a safeguarded version: ensure sufficient decrease w.r.t. the worst value in previous M iterations.

Acceleration by Continuation

- IST/FBS/PGA can be very slow if τ is very small and/or f is poorly conditioned.
- A very simple acceleration strategy: **continuation/homotopy**

Initialization: Set $\tau_0 \gg \tau$, starting point $\bar{x}$, factor $\sigma \in (0, 1)$, and $k = 0$.

Iterations: Find approx solution $x(\tau_k)$ of $\min_x f(x) + \tau_k \psi(x)$, starting from $\bar{x}$;

if $\tau_k = \tau_f$ **STOP**;

Set $\tau_{k+1} \leftarrow \max(\tau_f, \sigma \tau_k)$ and $\bar{x} \leftarrow x(\tau_k)$;

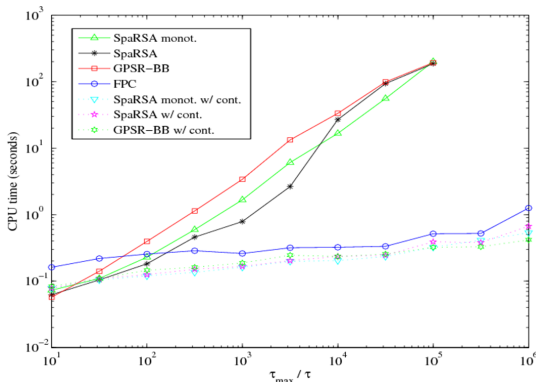
- Often the solution path $x(\tau)$, for a **range** of values of τ is desired, anyway (e.g., within an outer method to choose an optimal τ)
- Shown to be very effective in practice (Hale et al., 2008; Wright et al., 2009). Recently analyzed by Xiao and Zhang (2012).

Acceleration by Continuation: An Example

Classical **sparse reconstruction** problem (Wright et al., 2009)

$$\hat{x} \in \arg \min_x \frac{1}{2} \|Bx - b\|_2^2 + \tau \|x\|_1$$

with $B \in \mathbb{R}^{1024 \times 4096}$ (thus $x \in \mathbb{R}^{4096}$ and $b \in \mathbb{R}^{1024}$).



A Final Touch: Debiasing

Consider problems of the form $\hat{x} \in \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|Bx - b\|_2^2 + \tau \|x\|_1$

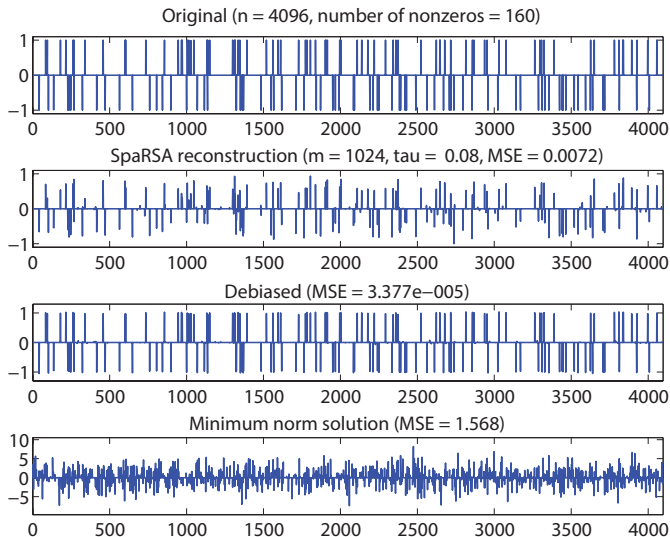
Often, the original goal was to minimize the quadratic term, after the support of x had been found. But the ℓ_1 term can cause the nonzero values of x_i to be “suppressed.”

Debiasing:

- ✓ find the zero set (complement of the support of $\hat{x}$):
 $\mathcal{Z}(\hat{x}) = \{1, \dots, n\} \setminus \text{supp}(\hat{x})$.
- ✓ solve $\min_x \|Bx - b\|_2^2$ s.t. $x_{\mathcal{Z}(\hat{x})} = 0$. (Fix the zeros and solve an unconstrained problem over the support.)

Often, this problem has to be solved using an algorithm that only involves products by B and B^T , since this matrix cannot be partitioned.

Effect of Debiasing



Example: Matrix Recovery (Toh and Yun, 2010)

$$\widehat{M} \in \arg \min_{M \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\Phi(M) - U\|_F^2 + \mu \|M\|_*$$

The proximal algorithm (IST) is as before:

linear operator

...its adjoint

$$X_{k+1} = \text{svt}_{\mu \beta_k} \left(X_k - \beta_k \Phi^*(\Phi(X_k) - U) \right)$$

Matrix completion: $\Phi(X) = X_\Omega$ (subset of entries) $|\Omega| = p$

Unknown M			IST				APG (FISTA)		
n/r	p	p/d_r	μ	iter	#sv	error	iter	#sv	error
100/10	5666	3	8.21e-03	7723	61	1.88e-01	655	13	1.06e-03
200/10	15665	4	1.05e-02	12180	96	2.45e-01	812	12	1.02e-03
500/10	49471	5	1.21e-02	10900	203	5.91e-01	1132	16	7.63e-04

Unknown M			continuation				APG + continuation		
n/r	p	p/d_r	μ	iter	#sv	error	iter	#sv	error
100/10	5666	3	8.21e-03	429	32	1.06e-03	74	10	1.46e-04
200/10	15665	4	1.05e-02	278	49	4.38e-04	73	10	1.02e-04
500/10	49471	5	1.21e-02	484	125	5.50e-04	72	10	8.06e-05

...the importance of acceleration!

Conditional Gradient

Also known as “Frank-Wolfe” after the authors who devised it in the 1950s. Later analysis by Dunn (around 1990). Suddenly a topic of enormous renewed interest; see for example (Jaggi, 2012).

$$\min_{x \in \Omega} f(x),$$

where f is a convex function and Ω is a closed, bounded, convex set.

Start at $x_0 \in \Omega$. At iteration k :

$$v_k := \arg \min_{v \in \Omega} v^T \nabla f(x_k);$$

$$x_{k+1} := x_k + \alpha_k(v_k - x_k), \quad \alpha_k = \frac{2}{k+2}.$$

- Potentially useful when it is easy to minimize a linear function over the *original* constraint set Ω ;
- Admits an elementary convergence theory: $1/k$ sublinear rate.
- Same convergence theory holds if we use a line search for α_k .

Conditional Gradient Convergence

Diameter of Ω is $D := \max_{x,y \in \Omega} \|x - y\|$.

Theorem

Suppose that f is convex, ∇f has Lipschitz L , Ω is closed, bounded, convex with diameter D . Then conditional gradient with $\alpha_k = 2/(k+2)$ yields

$$f(x^k) - f(x^*) \leq \frac{2LD^2}{k+2}, \quad k = 1, 2, \dots$$

Proof. Setting $x = x^k$ and $y = x^{k+1} = x^k + \alpha_k(v^k - x^k)$ in the usual bound, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T (v^k - x^k) + \frac{1}{2} \alpha_k^2 L \|v^k - x^k\|^2 \\ &\leq f(x^k) + \alpha_k \nabla f(x^k)^T (v^k - x^k) + \frac{1}{2} \alpha_k^2 L D^2, \end{aligned} \quad (12)$$

where the second inequality comes from the definition of D .

Condition Gradient Convergence, continued

For the first-order term, we have

$$\nabla f(x^k)^T(v^k - x^k) \leq \nabla f(x^k)^T(x^* - x^k) \leq f(x^*) - f(x^k).$$

Substitute in (12) and subtract $f(x^*)$ from both sides:

$$f(x^{k+1}) - f(x^*) \leq (1 - \alpha_k)[f(x^k) - f(x^*)] + \frac{1}{2}\alpha_k^2 LD^2.$$

Now **Induction**. For $k = 0$, with $\alpha_0 = 1$, have

$$f(x^1) - f(x^*) \leq \frac{1}{2}LD^2 < \frac{2}{3}LD^2,$$

as required. Suppose the claim holds for k , and prove for $k + 1$. We have

...

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &\leq \left(1 - \frac{2}{k+2}\right) [f(x^k) - f(x^*)] + \frac{1}{2} \frac{4}{(k+2)^2} LD^2 \\
&= LD^2 \left[\frac{2k}{(k+2)^2} + \frac{2}{(k+2)^2} \right] \\
&= 2LD^2 \frac{(k+1)}{(k+2)^2} \\
&= 2LD^2 \frac{k+1}{k+2} \frac{1}{k+2} \\
&\leq 2LD^2 \frac{k+2}{k+3} \frac{1}{k+2} = \frac{2LD^2}{k+3},
\end{aligned}$$

as required.

Conditional Gradient for Atomic-Norm Constraints

Conditional Gradient is particularly useful for optimization over atomic-norm constraints.

$$\min f(x) \text{ s.t. } \|x\|_{\mathcal{A}} \leq \tau.$$

Reminder: Given the set of atoms $\mathcal{A}$ (possibly infinite) we have

$$\|x\|_{\mathcal{A}} := \inf \left\{ \sum_{a \in \mathcal{A}} c_a : x = \sum_{a \in \mathcal{A}} c_a a, c_a \geq 0 \right\}.$$

The search direction v_k is $\tau \bar{a}_k$, where

$$\bar{a}_k := \arg \min_{a \in \mathcal{A}} \langle a, \nabla f(x_k) \rangle.$$

That is, we seek the atom that lines up best with the negative gradient direction $-\nabla f(x_k)$.

Generating Atoms

We can think of each step as the “addition of a new atom to the basis.” Note that x_k is expressed in terms of $\{\bar{a}_0, \bar{a}_1, \dots, \bar{a}_k\}$.

If few iterations are needed to find a solution of acceptable accuracy, then we have an approximate solution that's represented in terms of few atoms, that is, **sparse** or compactly represented.

For many atomic sets $\mathcal{A}$ of interest, the new atom can be found cheaply.

Example: For the constraint $\|x\|_1 \leq \tau$, the atoms are $\{\pm e_i : i = 1, 2, \dots, n\}$. if i_k is the index at which $|\nabla f(x_k)|_i$ attains its maximum, we have

$$\bar{a}_k = -\text{sign}([\nabla f(x_k)]_{i_k}) e_{i_k}$$

Example: For the constraint $\|x\|_\infty \leq \tau$, the atoms are the 2^n vectors with entries ± 1 . We have

$$[\bar{a}_k]_i = -\text{sign}[\nabla f(x_k)]_i, \quad i = 1, 2, \dots, n.$$

More Examples

Example: Nuclear Norm. For the constraint $\|X\|_* \leq \tau$, for which the atoms are the rank-one matrices, we have $\bar{A}_k = u_k v_k^T$, where u_k and v_k are the first columns of the matrices U_k and V_k obtained from the SVD $\nabla f(X_k) = U_k \Sigma_k V_k^T$.

Example: sum-of- ℓ_2 . For the constraint

$$\sum_{i=1}^m \|x_{[i]}\|_2 \leq \tau,$$

the atoms are the vectors a that contain all zeros except for a vector $u_{[i]}$ with unit 2-norm in the $[i]$ block position. (Infinitely many.) The atom $\bar{a}_k$ contains nonzero components in the block i_k for which $\|[\nabla f(x_k)]_{[i]}\|$ is maximized, and the nonzero part is

$$u_{[i]} = -[\nabla f(x_k)]_{[i_k]} / \|[\nabla f(x_k)]_{[i_k]}\|.$$

Reoptimizing. Instead of fixing the contribution α_k from each atom at the time it joins the basis, we can periodically and approximately reoptimize over the current basis.

- This is a finite dimension optimization problem over the (nonnegative) coefficients of the basis atoms.
- It need only be solved approximately.
- If any coefficient is reduced to zero, it can be dropped from the basis.

Dropping Atoms. Sparsity of the solution can be improved by dropping atoms from the basis, if doing so does not degrade the value of f too much (see (Rao et al., 2013)).

In the important least-squares case, the effect of dropping can be evaluated efficiently.

- Akaike, H. (1959). On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Annals of the Institute of Statistics and Mathematics of Tokyo*, 11:1–17.
- Barzilai, J. and Borwein, J. (1988). Two point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148.
- Beck, A. and Teboulle, M. (2009a). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.
- Beck, A. and Teboulle, M. (2009b). A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.
- Bioucas-Dias, J. and Figueiredo, M. (2007). A new twist: two-step iterativeshrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16:2992–3004.
- Brucker, P. (1984). An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3:163–166.
- Combettes, P. and Pesquet, J.-C. (2011). Signal recovery by proximal forward-backward splitting. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer.
- Hale, E., Yin, W., and Zhang, Y. (2008). Fixed-point continuation for l_1 -minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19:1107–1130.

References II

- Jaggi, M. (2012). Revisiting frank-wolfe: Projection-free sparse convex optimization. Ecole Polytechnique, France.
- Maculan, N. and de Paula, G. G. (1989). A linear-time median-finding algorithm for projecting a vector on the simplex of $\mathbb{R}^n$. *Operations Research Letters*, 8:219–222.
- Moreau, J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899.
- Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Doklady*, 27:372–376.
- Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York.
- Rao, N., Shah, P., Wright, S. J., and Nowak, R. (2013). A greedy forward-backward algorithm for atomic norm constrained minimization. In *Proceedings of ICASSP*.
- Toh, K.-C. and Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6:615–640.
- Wright, S., Nowak, R., and Figueiredo, M. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57:2479–2493.
- Xiao, L. and Zhang, T. (2012). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*. (to appear; available at <http://arxiv.org/abs/1203.3002>).