First-Order Methods

Stephen J. Wright[1]

[2]Computer Sciences Department,
University of Wisconsin-Madison.

IMA, August 2016

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 $\{x_k\}$, $k = 0, 1, 2, \ldots$ from

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

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 neighnorhood of such points.

Steepest Descent

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

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

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

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

Summing over first $T-1$ iterates ($k = 0, 1, \ldots, T-1$) 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 $\nabla f(x_k) \to 0$ if f is bounded below.

Convergence Rates

From the sum above we have that

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

and so

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

"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

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

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

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

Continued...

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

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

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

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

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

$$\begin{aligned} \phi_k &\leq \frac{C_1}{\sqrt{k}}, \ \frac{C_2}{k}, \ \frac{C_3}{k^2} && \text{(sublinear)} \\ \phi_{k+1} &\leq (1-c)\phi_k \ \text{ for some } c \in (0,1) && \text{(linear)} \\ \phi_{k+1} &= o(\phi_k) && \text{(superlinear).} \end{aligned}$$

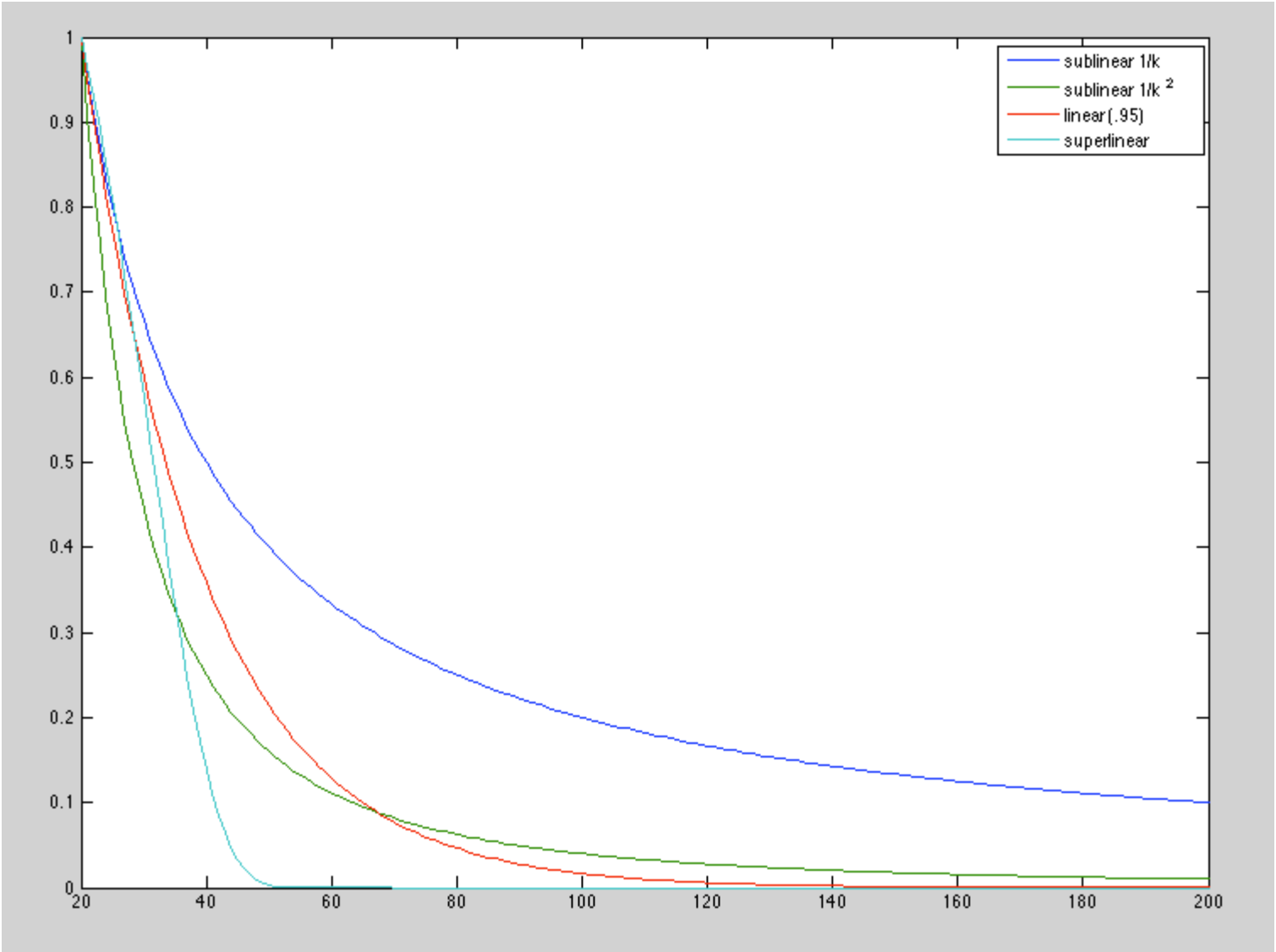
To achieve $\phi_T \leq \epsilon$ for some small positive tolerance ϵ, need

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

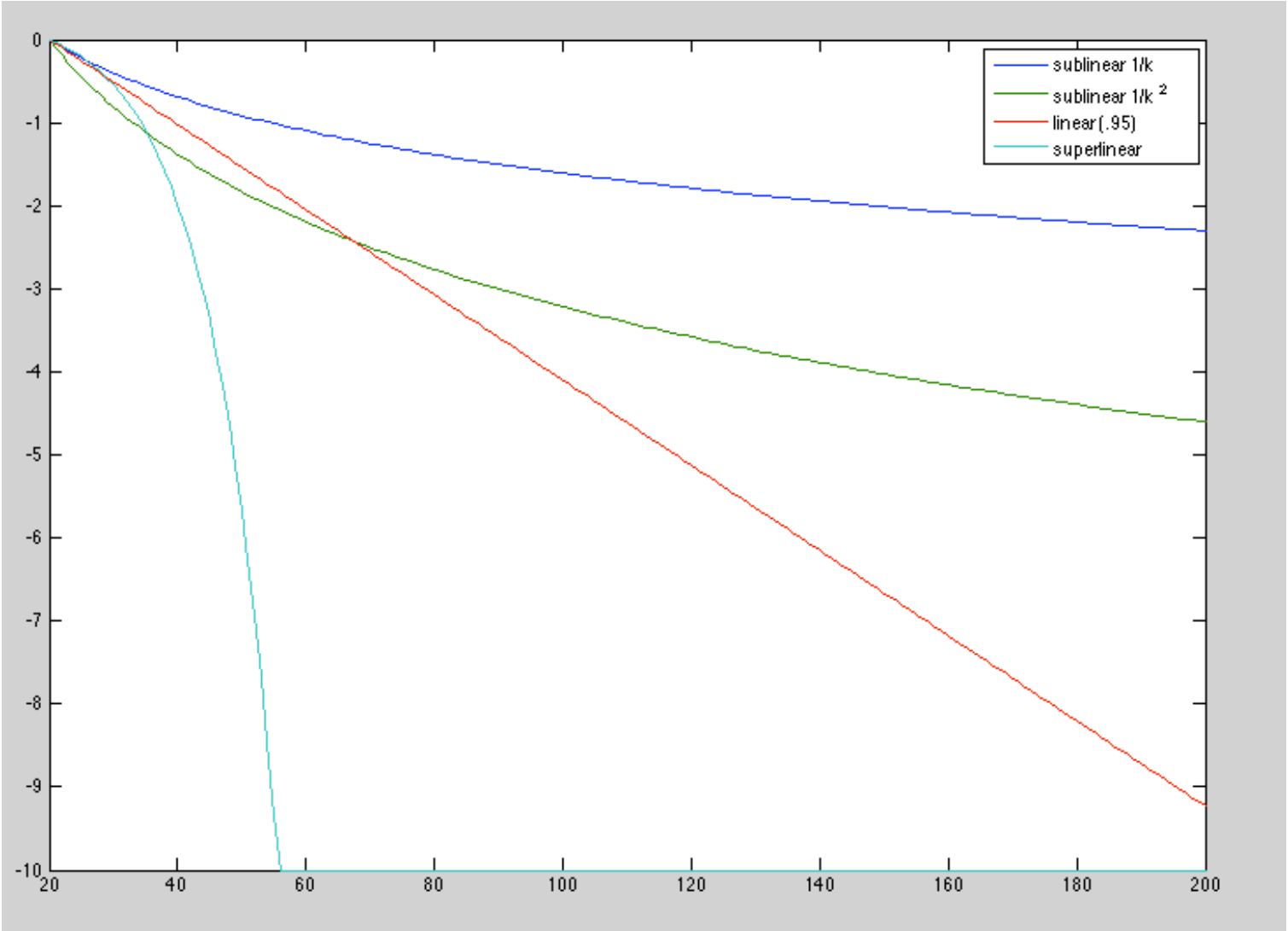
$$T = O\left(\frac{1}{c}\log\epsilon\right), \quad \text{for linear rate.}$$

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:

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

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

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.

More on KL

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

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

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

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

KI also holds when $f(x) = \sum_{i=1}^m h(a_i^T x)$, where $h : \mathbb{R} \to \mathbb{R}$ 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 $x_{k+1} \in x_0 + \text{span}(\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k))$.

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

and set $f(x) = (1/2)x^T Ax - e_1^T x$. 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 $i = k+1, k+2, \ldots, n$. 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

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

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

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

We impose weak Wolfe conditions on steplength α_k:

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

$$\nabla f(x_k + \alpha d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k \,. \tag{5b}$$

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

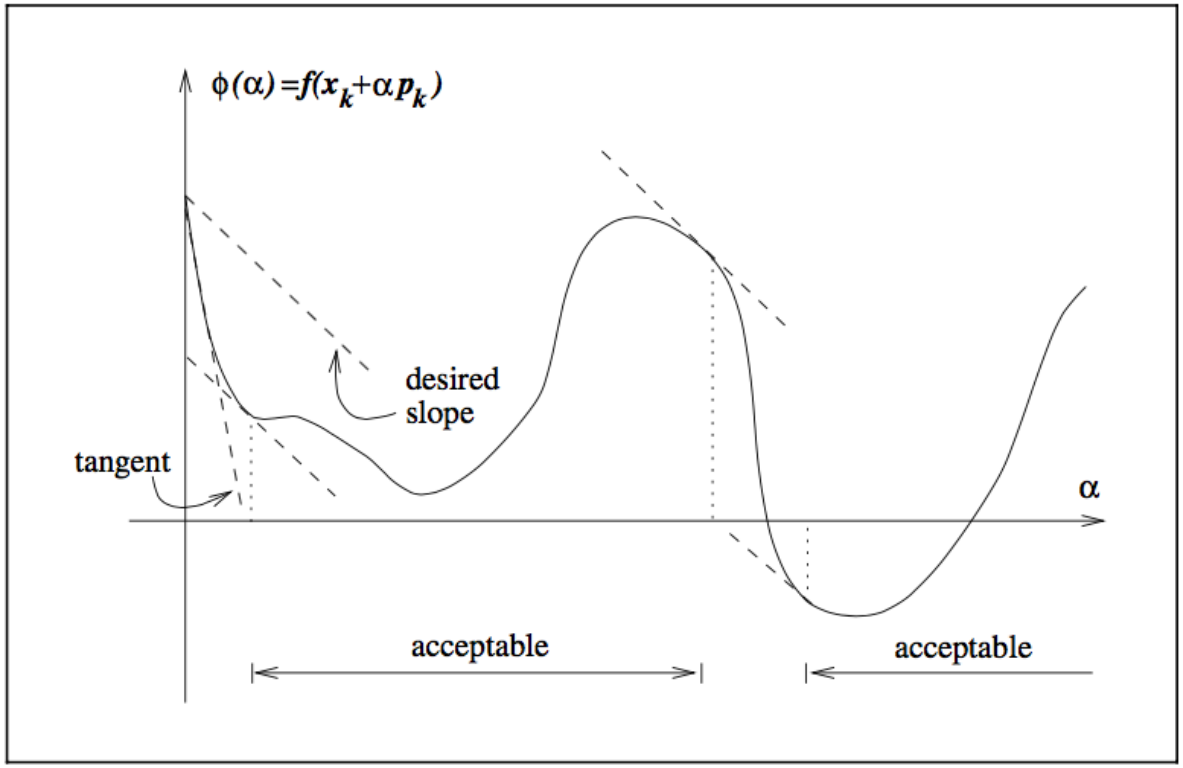


Figure 3.4 The curvature condition.

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 $\alpha_k = \bar{\alpha}, \frac{\bar{\alpha}}{2}, \frac{\bar{\alpha}}{4}, \frac{\bar{\alpha}}{8}, \ldots$ 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 or quadratic growth holds, we have

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

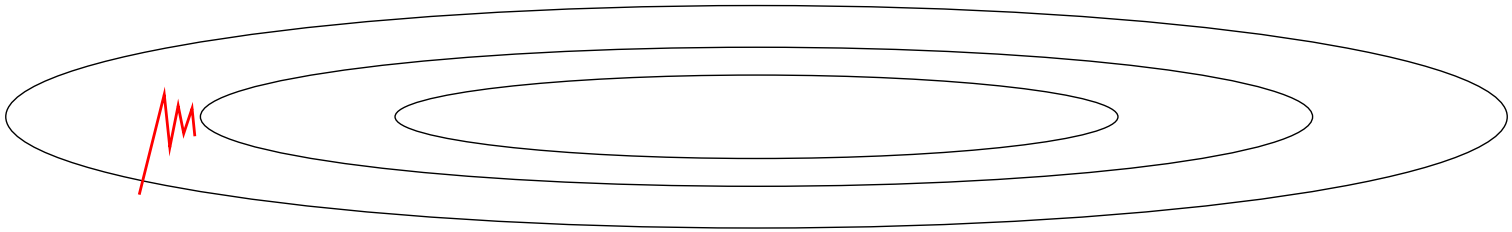
so that

$$\text{dist}(x, \text{solution set}) \leq \sqrt{\frac{1}{a_3}(f(x_k) - f^*)}.$$

So we can derive convergence rates on $\text{dist}(x, \text{solution set})$ from those of $f(x_k) - f^*$.

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 and Nesterov

Heavy Ball:

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

Nesterov's optimal method:

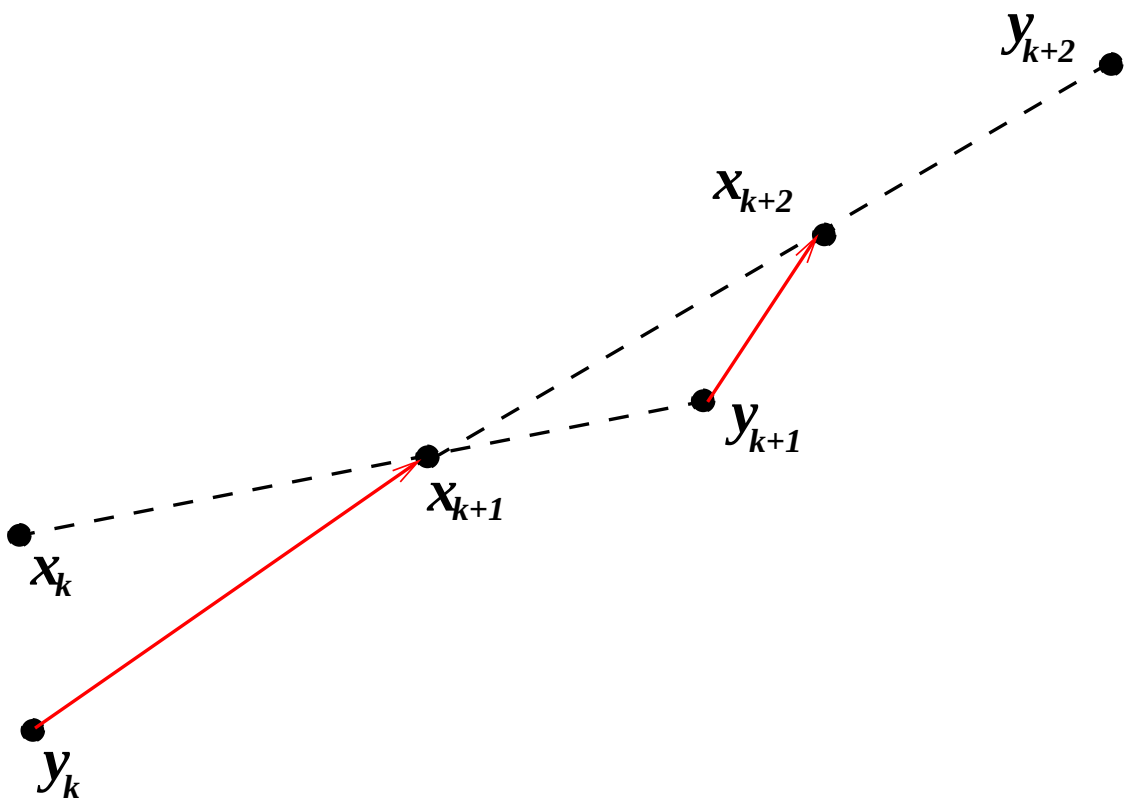
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k + \beta_k(x_k - x_{k-1})) + \beta_k(x_k - x_{k-1}).$$

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

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

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

Nesterov, illustrated



Separates the "gradient descent" and "momentum" step components.

Accelerated Gradient Convergence

Typical convergence:

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

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

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:

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

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

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

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

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

Multistep Methods: The Heavy-Ball

Matrix T has same eigenvalues as

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

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:

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

All eigenvalues have magnitude β!

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

The recurrence becomes

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

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:

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

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

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

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

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

Conjugate Gradient

Basic conjugate gradient (CG) step is

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

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

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

- Choose α_k to (approximately) miminize 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 Methods

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

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

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

FISTA

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:
$$x_k \leftarrow y_k - \tfrac{1}{L}\nabla f(y_k);$$
$$t_{k+1} \leftarrow \tfrac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right);$$
$$y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).$$

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.

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

where

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

Explicitly, we have

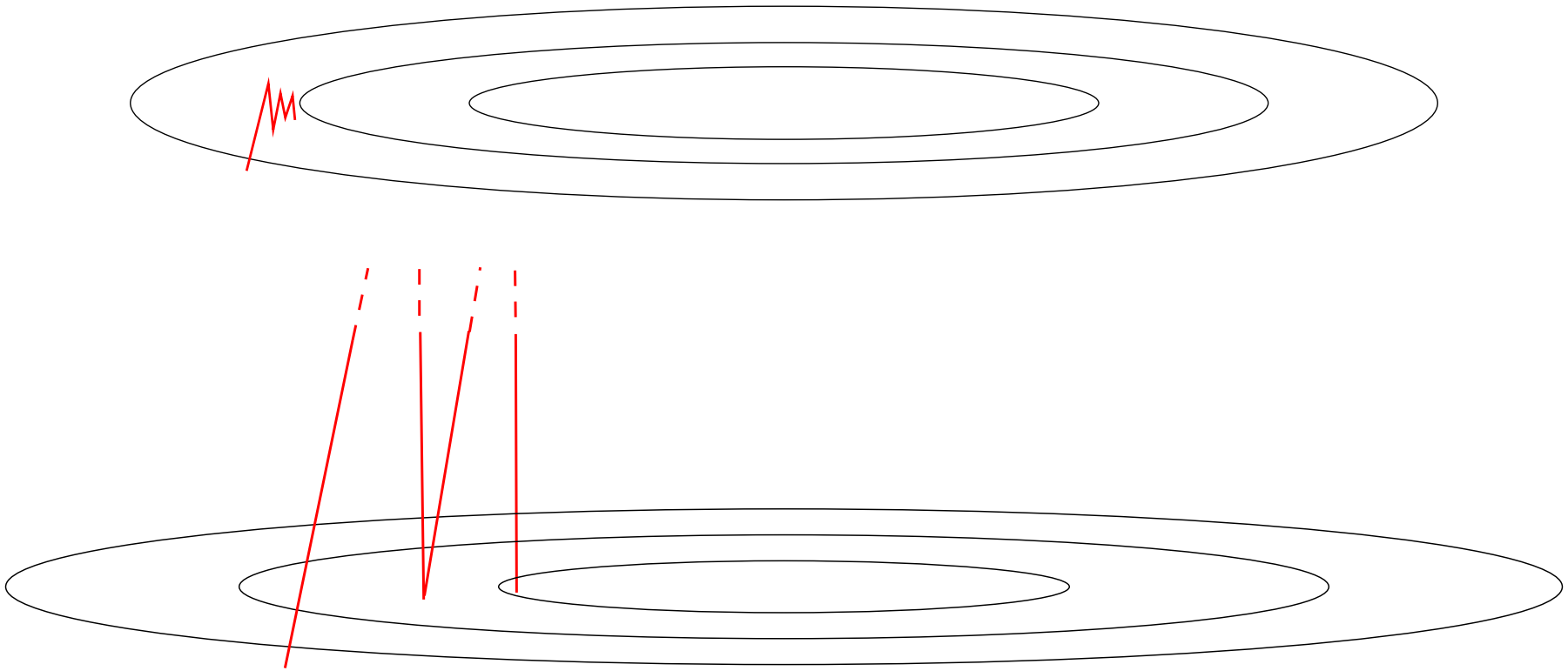
$$\alpha_k = \frac{s_k^T z_k}{z_k^T z_k}.$$

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 $x \in \Omega$ explicity 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.

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

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

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

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T (v^k - x^k) + \frac{1}{2}\alpha_k^2 L \|v^k - x^k\|^2 \\ &\leq f(x^k) + \alpha_k \nabla f(x^k)^T (v^k - x^k) + \frac{1}{2}\alpha_k^2 L D^2, \end{aligned} \tag{6}$$

where the second inequality comes from the definition of D.

Conditional 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 (6) 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.

References I

Akaike, H. (1959). On a successive transformation of probability distribution and its application to the analysis fo 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.

Jaggi, M. (2012). Revisiting frank-wolfe: Projection-free sparse convex optimization. Ecole Polytechnique, France.

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*.