1 The Problem

We are interested in unconstrained minimization of $f : \mathbb{R}^n \to \mathbb{R}$ given first order oracle

$$\text{minimize}_{x \in \mathbb{R}^n} \; f(x)$$

We may impose additional assumption on f, i.e. convex, L-lipschitz, μ-strongly convex

2 Gradient Descent

Gradient descent updates according to

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

for some stepsize $\alpha_k \geq 0$. Note $\alpha_k = \frac{1}{L}$ is the optimal stepsize.

2.1 Barzilai & Borwein Stepsize

A particular choice of stepsize that relaxes the constraint on monotonic descent is given by Barzilai & Borwein [1]. The idea is to choose α_k such that $\alpha_k g^k$ approximates the Newton update.

$$\alpha_k = \frac{\langle u^k, v^k \rangle}{\|v^k\|^2} \qquad \text{or} \qquad \alpha_k = \frac{\|u^k\|^2}{\langle u^k, v^k \rangle}$$

where

$$u^k = x^k - x^{k-1} \qquad v^k = \nabla f(x^k) - \nabla f(x^{k-1})$$

3 Nesterov's Accelerated Gradient

Nesterov's accelerated gradient achieves lower bound for minimization of function $f \in \mathscr{S}^1_{L,\mu}$ and improves the rate for gradient descent from $\mathcal{O}(\kappa \log \frac{1}{\epsilon})$ to $\mathcal{O}(\sqrt{\kappa} \log \frac{1}{\epsilon})$. Similarly, acceleration improves convergence rate for function $f \in \mathscr{F}^1_L$ from $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$.

3.1 Intuition

The following comes from Nesterov's book [2] and lecture note.

Definition. *A pair of sequences* $(\{\phi_k(x)\}_{k=0}^{\infty}, \{\lambda_k\}_{k=0}^{\infty})$ *where* $\lambda_k \geq 0$ *are called the estimating sequences of the function* $f(\cdot)$ *if*

1. $\lambda_k \to 0$ *and*
2. ***(lower bound)*** *for any* $x \in \mathbb{R}^n$ *and for all* $k \geq 0$, $\phi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \phi_0(x)$

In addition, If we can find some sequence of points $\{x^k\}_{k=0}^{\infty}$ such that

3. ***(upper bound)*** *for any* $x \in \mathbb{R}^n$, $f(x^k) \leq \phi_k(x)$

then the rate of convergence can be derived from convergence rate of λ_k, i.e.

$$f(x^k) - f^* \leq \lambda_k \left\{\phi_0^* - f^*\right\} \to 0$$

where $\phi_k^* = \min_{x\in\mathbb{R}^n} \phi_k(x)$. Intuitively, $\phi_k(\cdot)$ are approximations for $f(\cdot)$, providing tighter and tighter bound on the optimality gap $f(x^k) - f^*$ as $\lambda_k \to 0$. In addition, from (2) and (3), we have that the sequence $\left\{x^k\right\}$ converges to the minimizer of f.

$$f(x^k) \leq \phi_k(x^*) \leq f(x^*)$$

In [2], Nesterov showed that for $f \in \mathscr{S}^1_{\mu,L}$, we can construct estimating sequences for f recursively

$$\begin{aligned}
\lambda_{k+1} &= (1-\alpha_k)\lambda_k \\
\phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k L_k(x) \\
&\text{where} \quad L_k(x) = f(y^k) + \left\langle \nabla f(y^k), x - y^k \right\rangle + \frac{\mu}{2}\left\|x - y^k\right\|^2
\end{aligned}$$

where $\left\{y^k\right\}_{k=0}^{\infty}$ is an arbitrary sequence of points, coefficients $\{\alpha_k\}_{k=0}^{\infty}$ satisfy $\alpha_k \in (0,1)$ and $\sum_k \alpha_k = \infty$ with $\lambda_0 = 1$ and that $\phi_0(\cdot)$ is an arbitrary convex function. Note that ϕ_k is simply a convex combination of the previous approximate ϕ_{k-1} and a quadratic lower bound L_{k-1} on f, at some carefully chosen point y^{k-1}. If we let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\left\|x - v_0\right\|^2$ be a quadratic function, then $\phi_k(\cdot)$ has a convenient closed form expression

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}\left\|x - v_k\right\|^2$$

where $\left\{\gamma_k\right\}, \left\{v_k\right\}, \left\{\phi_k^*\right\}$ follow certain recurrence relation detailed in [2]. Additional constraint needs to be satisfied to ensure (3) holds.

1. For (3) to hold, it must be that $f(y^k) - \frac{1}{2L}\left\|\nabla f(y^k)\right\|^2 \geq f(x^{k+1})$, which can be achieved if we obtain x^{k+1} by taking a gradient step $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$ at y^k and apply descent lemma.
2. To apply the previous, we need the coefficient before $\left\|\nabla f(y^k)\right\|^2$ to agree, i.e. want α_k such that $L\alpha_k^2 = (a - \alpha_k)\gamma_k + \alpha_k\mu$.
3. Choose y^k accordingly to ensure (3) holds

By making these constraints invariant to iterative updates, we arrive at the accelerated gradient methods. In addition to the algebra tricks, there are efforts that tries to interpret what Nesterov's method is doing under the hoold. For example, [3] interpreted Nesterov's accelerated method as a linear coupling of gradient descent and mirror descent . [4] showed that in the limit of small stepsizes (when taking the gradient step to obtain x^{k+1}) is equivalent to the dynamics of some continous second-order ODE.

3.2 Algorithm

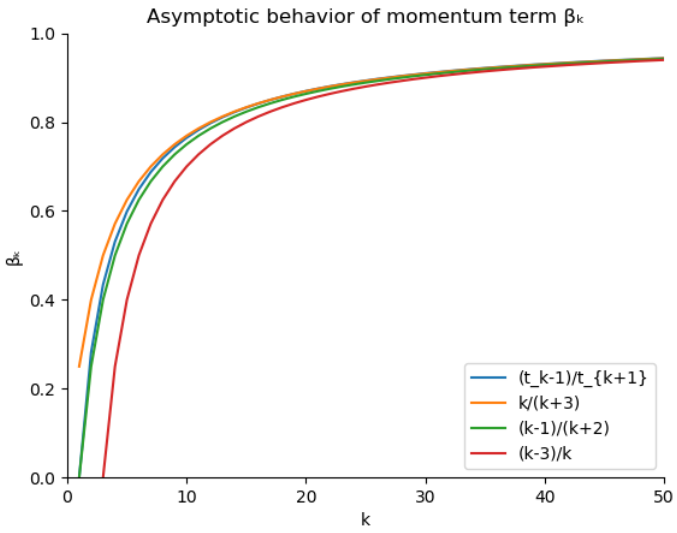
There are several equivalent algorithm for Nesterov's Accelerated Gradient Method. The following came from the original paper by Nesterov in 1983 [5] and later adapted to LASSO [6]. Assume $f \in \mathscr{F}_L^1$ is in the class of convex, L-Lipschitz continuous functions. Given $t_1 = 1$ and $y_1 = x_0$, accelerated gradient updates according to

$$\begin{aligned}
x^{k+1} &= y^k - \frac{1}{L}\nabla f(y^k) \\
t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\
y^{k+1} &= x^{k+1} + \frac{t_k - 1}{t_{k+1}}\left(x^{k+1} - x^k\right)
\end{aligned}$$

We can simplify the expression by noting that (slides)

$$\beta_k = \frac{t_k - 1}{t_{k+1}} = 1 - \frac{3}{k} + o(\frac{1}{k}) = \frac{k-3}{k} + o(\frac{1}{k})$$

The momentum coefficient is asymptotically equivalent to $\frac{k}{k+3}$ or $\frac{k-1}{k+2}$. The momentum term $\frac{k-1}{k+2}$ seems to be a good choice, since $\beta_1 = \frac{t_1 - 1}{t_2} = 0$.



And updates is now given by

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \frac{k-1}{k+2}\left(x^{k+1} - x^k\right)$$

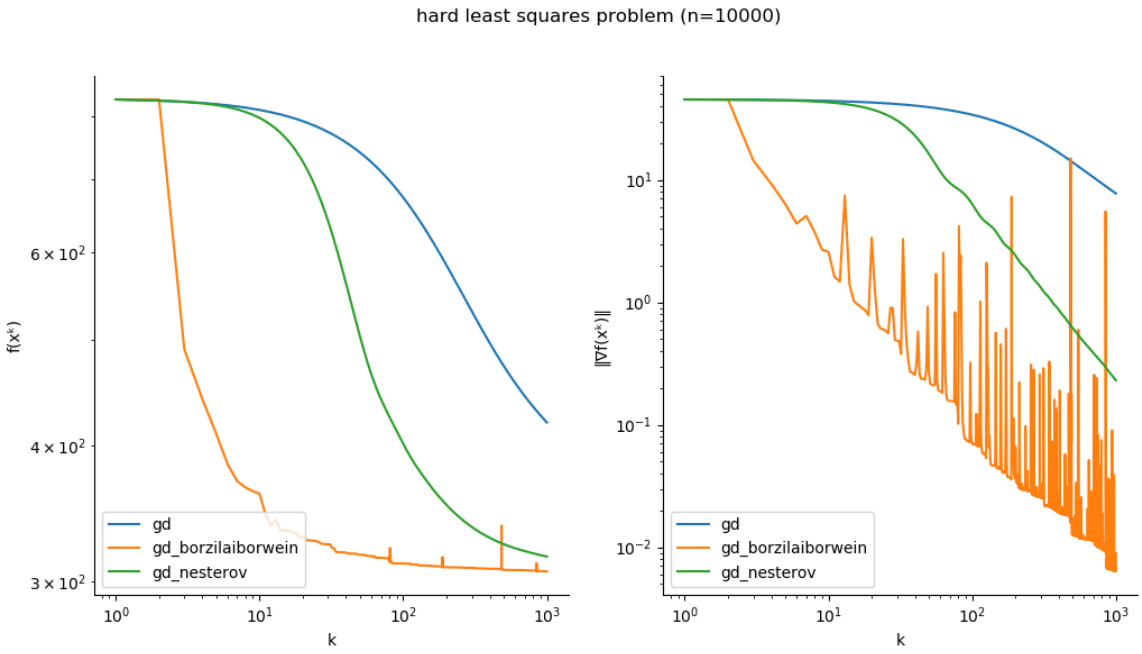
Another formulation of the algorithm comes from Nesterov's textbook [2]. If we take a constant step, i.e. $\frac{1}{L}$, to find the x^{k+1}, and that we pick $\alpha_0 = \sqrt{\frac{\mu}{L}} = 1/\sqrt{\kappa}$, which is the interpolating coefficient for recursive construction of the estimating sequence. Then we have the follwing updates

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(x^{k+1} - x^k)$$

However, in practice the condition number κ is hard to compute.

4 Numerical Experiments

We are given a hard least squares problem of minimizing $f(x) = \frac{1}{2}\left\|D^T x - b\right\|_2^2$ where $D \in \mathbb{R}^{n\times(n+1)}$ is the differencing matrix, with all -1 on the main diagonal and all 1 on the superdiagonal. The gradient is given by $\nabla f(x) = D(D^T x - b)$. We compare gradient descent with either constant stepsize or using barzilai borwein stepsize, and nesterov's accelerated gradient descent.



We see that the barzilai borwein stepsize is the fastest method, followed by nesterov's accelerated gradient, then the naive gradient descent method.

References

[1] Jonathan Barzilai and Jonathan M. Borwein. "Two-Point Step Size Gradient Methods". In: *IMA Journal of Numerical Analysis* 8.1 (Jan. 1, 1988). Publisher: Oxford Academic, pp. 141–148. ISSN: 0272-4979. DOI: 10.1093/imanum/8.1.141. URL: https://academic.oup.com/imajna/article/8/1/141/802460 (visited on 03/25/2020).

[2] Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. 2004. ISBN: 978-1-4020-7553-7. URL: https://dial.uclouvain.be/pr/boreal/object/boreal:116858 (visited on 03/27/2020).

[3] Zeyuan Allen-Zhu and Lorenzo Orecchia. "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent". In: *arXiv:1407.1537 [cs, math, stat]* (Nov. 7, 2016). arXiv: 1407.1537. URL: http://arxiv.org/abs/1407.1537 (visited on 03/24/2020).

[4] Weijie Su, Stephen Boyd, and Emmanuel J. Candes. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *arXiv:1503.01243 [math, stat]* (Oct. 27, 2015). arXiv: 1503.01243. URL: http://arxiv.org/abs/1503.01243 (visited on 03/28/2020).

[5] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate O(1/k2)." In: *Soviet Mathematics Doklady* 27 ((2) 1983), pp. 372–376.

[6] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (Jan. 2009), pp. 183–202. ISSN: 1936-4954. DOI: 10.1137/080716542. URL: http://epubs.siam.org/doi/10.1137/080716542 (visited on 03/27/2020).