1 Smooth Convex Optimization

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

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

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

1.1 Gradient Descent

Gradient descent achieves sublinear convergence $\mathcal{O}(\frac{1}{\epsilon})$ for $f \in \mathscr{F}_L^1$ and $\mathcal{O}(\log \frac{1}{\epsilon})$ for $f \in \mathscr{S}_{L,\mu}^1$.

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

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

1.2 Gradient Descent with Barzilai & Borwein Stepsize

Barzilai & Borwein stepsize relaxes the constraint on monotonic descent [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

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

This algorithm enjoys fast empirical convergence.

1.3 Nesterov's Accelerated Gradient

Nesterov's accelerated gradient achieves lower bound for minimization of function $f \in \mathscr{S}_{L,\mu}^1$ 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 $f \in \mathscr{F}_L^1$ from $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$.

1.3.1 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 $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2$ be a quadratic function, then $\phi_k(\cdot)$ has a convenient closed form expression

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

where $\{\gamma_k\}, \{v_k\}, \{\phi_k^*\}$ 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.

1.3.2 The Algorithm

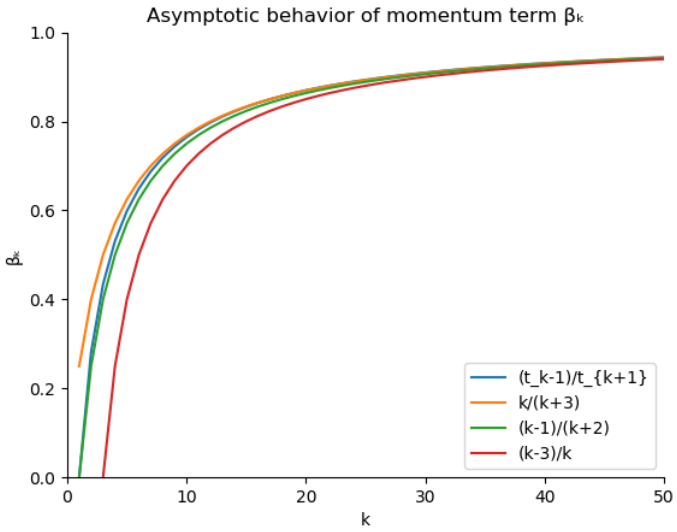
There are several equivalent algorithm for Nesterov's Accelerated Gradient Method. The following came from the original paper by Nesterov in 1983 [5] and later adapted to LASSO [6]. Assume $f \in \mathscr{F}^1_L$. Given $t_1 = 1$ and $y_1 = x_0$, accelerated gradient updates according to

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

We can simplify the expression by noting that (slides)

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

The momentum coefficient is asymptotically equivalent to $\frac{k-1}{k+2}$ ($\frac{t_1-1}{t_2} = 0$)



And updates is now given by

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

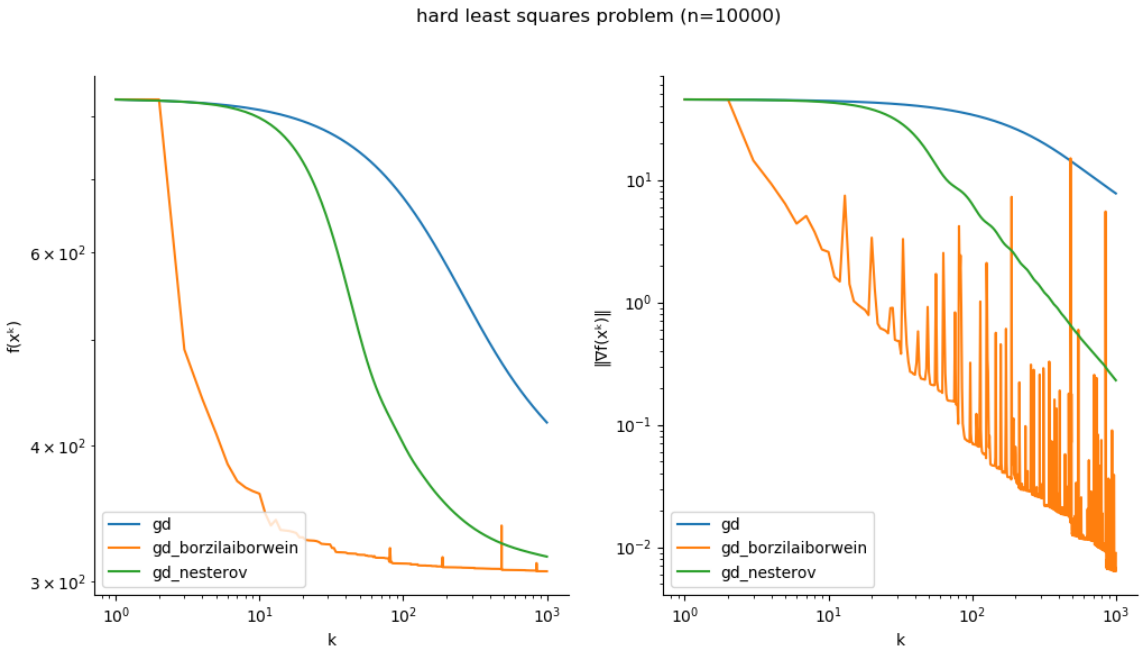
Another formulation of the algorithm comes from Nesterov's textbook [2]. If we take a constant step, i.e. $\frac{1}{L}$, to find the x^{k+1}, and that we pick $\alpha_0 = \sqrt{\frac{\mu}{L}} = 1/\sqrt{\kappa}$, which is the interpolating coefficient for recursive construction of the estimating sequence.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

1.4 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.

2 Nonsmooth Convex Optimization

We are interested in unconstrained minimization of convex, possibly nondifferentiable, $f : \mathbb{R}^n \to \mathbb{R}$

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

given first order oracle

2.1 Subgradient Method

Given bounded subgradient $\left\|g^k\right\| \leq G$ and bounded domain $\left\|x^0 - x^*\right\| \leq R$, subgradient method is in a sense optimal as it achieves the lower bound $\mathcal{O}(\frac{1}{\epsilon^2})$ for this problem class. Subgradient method iteratively updates as follows

$$x^{k+1} = x^k - \alpha_k g^k$$

where $g^k \in \partial f(x^k)$ is *any* subgradient of f. First order optimality condition is now $0 \in \partial f(x^*)$, which is impossible to test for nontrivial function f. Therefore, using $\left\|g^k\right\| \leq \epsilon$ is not informative and subgradient method does not really have a stopping criterion.

2.1.1 Solving Support Vector Machine w/ Subgradient Method

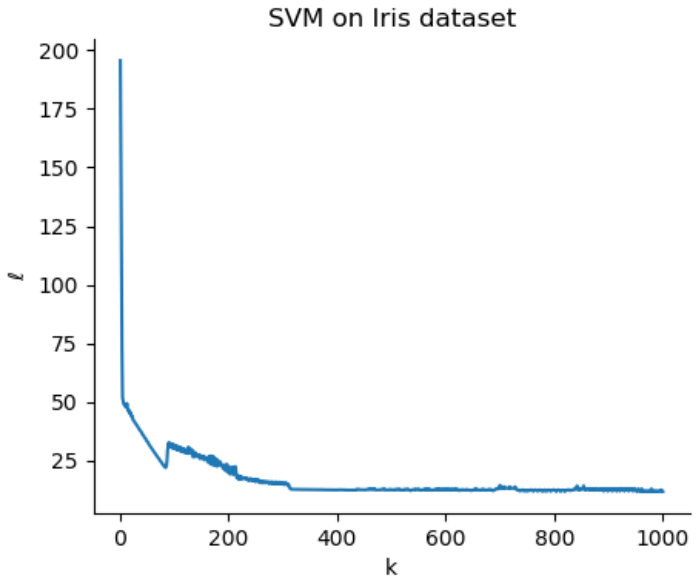
We are given data $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n \; y_i \in \{\pm 1\}\}$, support vector machine is supervised learning model that tries to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that the empirical risk and regularizer on w is minimized

$$\text{minimize}_{w,b} \quad \frac{1}{2}\|w\|_2^2 + \lambda \sum_{i=1}^{m} \max\left[0, 1 - y_i(w^T x_i + b)\right] \quad (:= f(w,b))$$

Support vector machines can be solved using subgradient method. We first find a subgradient of f

$$g_w^k = w^k - \lambda \sum_{i \in [m]: y_i(w^T x_i + b) < 1} y_i x_i$$
$$g_b = -\lambda \sum_{i \in [m]: y_i(w^T x_i + b) < 1} y_i$$

where we haved picked $0 \in \partial(\max 0, 1 - y_i(w^T x_i + b))$ when $y_i(w^T x_i + b) = 1$, the only case where the *max term* is non-differentiable. When tested on the Iris dataset, subgradient method worked!



3 Stochastic Optimization

We are interested in constrained minimization of $f : \mathbb{R}^n \to \mathbb{R}$

$$\text{minimize}_{x \in \mathcal{X}} \ [f(x) = \mathbb{E}[F(x, \xi)]]$$

where $\mathcal{X} \subset \mathbb{R}^n$ is closed, bounded convex set. ξ is a random variable, and $F(\cdot, \xi)$ is convex for all $\xi \in \Xi$, and therefore $f(\cdot)$ is convex. For uniform p_ξ over finite alphabets of size n, the problem reduces to finite sum problem (or sample average approximation (SAA))

$$\text{minimize}_{x \in \mathcal{X}} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right]$$

Assume we can

1. Sample $\xi_1, \xi_2, \cdots \overset{i.i.d.}{\sim} p_\xi$
2. Given $(x, \xi) \in \mathcal{X} \times \Xi$, a first order oracle that returns a subgradient vector $G(x, \xi) \in \partial_x F(x, \xi)$. We also assume that G is unbiased, i.e. $g(x) := \mathbb{E}[G(x, \xi)] \in \partial f(x)$

3.1 Stochastic Gradient Method

We can show that if $f \in \mathscr{S}^1_{L,\mu}$, the choice of $\alpha_k = \mathcal{O}(1/k)$ yields sublinear convergence of $\mathcal{O}(\frac{1}{\epsilon})$ for last iterates. If $f \in \mathscr{F}^1_L$, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ yields a sublinear convergence of $\mathcal{O}(\frac{1}{\epsilon^2})$ for average iterates. Stochastic gradient method (or Stochastic Approximation (SA) algorithms) solves the problem by

$$x^{k+1} = \Pi_{\mathcal{X}}\left(x^k - \alpha_k G(x^k, \xi_k)\right)$$

where $\alpha_k > 0$ are stepsizes, $\Pi_{\mathcal{X}}(\cdot)$ is the projection operator to convex set. It is important to note that the current iterate x^k are functions of random variables $x^k := x^k(\xi_{[k-1]})$ where $\xi_{[k-1]} = (\xi_1, \cdots, \xi_{k-1})$, and therefore are random variables themselves. In addition, $x^k \perp\!\!\!\perp \xi_k$.

3.2 Convergence

Derivations copied from [7], [8] and slides. We assume

1. bounded variance for stochastic subgradient, which translates to $\mathbb{E}_\xi\left[G(x,\xi)\right] \leq M^2$ given $x \in \mathcal{X}$.
2. bounded $\mathcal{X}$ where radius given by $D_{\mathcal{X}} = \max_{x\in\mathcal{X}} \|x - x^*\|_2$.

We outline implications of some assumptions

1. If f is convex, then

$$f(x') \geq f(x) + \left\langle g(x), x' - x\right\rangle \qquad \forall\, x, x' \in \mathcal{X} \tag{1}$$

2. If f has L lipschitz continuous gradients, then

$$\left\|\nabla f(x') - \nabla f(x)\right\| \leq L\left\|x' - x\right\| \qquad \forall\, x, x' \in \mathcal{X} \tag{2}$$

$$f(x) - f(x^*) \leq \frac{1}{2}L\left\|x - x^*\right\| \qquad \forall\, x \in \mathcal{X} \qquad \text{(descent lemma)}$$

3. If f is μ-strongly convex, then

$$\left\langle \nabla f(x') - \nabla f(x), x' - x\right\rangle \geq \mu\left\|x' - x\right\|_2^2 \qquad \forall\, x \in \mathcal{X} \tag{3}$$

$$\mu\left\|x - x^*\right\|^2 \leq \mu\left\langle g(x) - g(x^*), x - x^*\right\rangle = \left\langle g(x), x - x^*\right\rangle \qquad \forall\, x \in \mathcal{X},\ g(x) \in \partial f(x) \tag{4}$$

We first derive some preliminary results. Using iterated expecatation, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle G(x^k,\xi_k), x^k - x^*\right\rangle\right] &= \mathbb{E}_{\xi_{[k-1]}}\left[\mathbb{E}_{\xi_k}\left[\left\langle G(x^k(\xi_{[k-1]}),\xi_k), x^k(\xi_{[k-1]}) - x^*\right\rangle\right] \mid \xi_{[k-1]}\right] \\
&= \mathbb{E}_{\xi_{[k-1]}}\left[\left\langle \mathbb{E}_{\xi_k}\left[G(x^k(\xi_{[k-1]}),\xi_k) \mid \xi_{[k-1]}\right], x^k(\xi_{[k-1]}) - x^*\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle g(x^k), x^k - x^*\right\rangle\right]
\end{aligned} \tag{5}$$

where the expectation is taken w.r.t $\xi_{[k-1]}$. We first derive a bound on $R_k = \left\|x^k - x^*\right\|_2^2$ and $r_k = \mathbb{E}\left[R_k\right]$,

$$
\begin{aligned}
R_{k+1} &= \left\|x^k - x^*\right\|^2 \\
&= \left\|\Pi_{\mathcal{X}}\left(x^k - \alpha_k G(x^k, \xi_k)\right) - \Pi_{\mathcal{X}}(x^*)\right\| && (\Pi_{\mathcal{X}}(x^*) = x^*) \\
&\leq \left\|x^k - \alpha_k (G * x^k, \xi_k) - x^*\right\|^2 && (\text{nonexpansive of } \Pi(\cdot)\ \|\Pi_{\mathcal{X}}(x') - \Pi_{\mathcal{X}}(x)\| \leq \|x' - x\|) \\
&\leq R^k - 2\alpha_k \left\langle G(x^k, \xi_k), x^k - x^* \right\rangle + \alpha_k^2 \left\|G(x^k, \xi_k)\right\|^2 \\
r_{k+1} &\leq r_k - 2\alpha_k \mathbb{E}\left[\left\langle G(x^k, \xi_k), x^k - x^* \right\rangle\right] + \alpha_k^2 \mathbb{E}\left[\left\|G(x^k, \xi_k)\right\|^2\right] && (\text{Take expectation w.r.t. } \xi_{[k]}) \\
&= r_k - 2\alpha_k \mathbb{E}\left[\left\langle g(x^k), x^k - x^* \right\rangle\right] + \alpha_k^2 M^2 && (\text{By (5) and bounded variance})
\end{aligned}
$$

3.2.1 Strongly Convex Case

If $f \in \mathscr{S}_{L,\mu}^1$, using (4), we have

$$
r_{k+1} \leq r_k - 2\alpha_k \mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] + \alpha_k^2 M^2 = (1 - 2\mu\alpha_k) r_k + \alpha_k^2 M^2
$$

If we choose $\alpha_k = \theta/(k+1)$, where $\theta > 1/(2\mu)$. It could be shown by induction that [7]

$$
r_k \leq \frac{c_\theta}{k+1} \qquad \text{where} \qquad c_\theta = \max\left\{\frac{2\theta^2 M^2}{2\mu\theta - 1}, r_0\right\}
$$

By (descent lemma), we derive bound on the objective value

$$
\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq \frac{1}{2} L \mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] \leq \frac{L c_\theta}{2(k+1)}
$$

Therefore, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\epsilon})$ yields last iterate convergence rate of $\mathcal{O}(\frac{1}{\epsilon})$

3.2.2 Convex Case

[7] indicates that we need to increase the stepsize ($\mathcal{O}(\frac{1}{k})$ to $\mathcal{O}(\frac{1}{\sqrt{k}})$)) to ensure faster convergence rate for general convex problems, at the cost of *more noisy* trajectory. To suppress the noise, we use average iterates $\left\{x^k\right\}$ rather than last iterates as solution to the problem.

$$
\begin{aligned}
r_{k+1} &\leq r_k - 2\alpha_k \mathbb{E}\left[\left\langle g(x^k), x^k - x^* \right\rangle\right] + \alpha_k^2 M^2 \\
2\alpha_k \mathbb{E}\left[f(x^k) - f(x^*)\right] &\leq 2\alpha_k \mathbb{E}\left[\left\langle g(x^k), x^k - x^* \right\rangle\right] \leq r_k - r_{k+1} + \alpha_k^2 M^2 && (\text{Rearrange, and by 1}) \\
\sum_{i=1}^k \left(2\alpha_i \mathbb{E}\left[f(x^i) - f(x^*)\right]\right) &\leq \sum_{i=1}^k \left(r_i - r_{i+1} + \alpha_i M^2\right) = r_1 + \sum_{i=1}^k \alpha_i^2 && (\text{Telescope}) \\
\sum_{i=1}^k \gamma_i \mathbb{E}\left[(f(x^i) - f(x^*))\right] &= \mathbb{E}\left[\sum_{i=1}^k \gamma_i (f(x^i) - f(x^*))\right] \leq \frac{r_1 + M^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i} \\
& && (\text{Let } \gamma_i = \alpha_i / \textstyle\sum_i \alpha_i. \text{ Divide by } 2\textstyle\sum_i \alpha_i. \text{ Use linearity of expectation}) \\
\mathbb{E}\left[f(\tilde{x}^k) - f(x^*)\right] &\leq \frac{r_1 + M^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i} \\
& && (\text{Let } \tilde{x}^k = \textstyle\sum_{i=1}^k \gamma_i x^i.\ f(\tilde{x}^k) \leq \textstyle\sum_i \gamma_i f(x^i) \text{ by convexity of } f.\ \textstyle\sum_i \gamma_i = 1)
\end{aligned}
$$

We derive tighest bound by finding minimal value of $\alpha_k = \alpha$ of the bound.

$$\mathbb{E}\left[f(\tilde{x}^k) - f(x^*)\right] \leq \frac{D_{\mathcal{X}} M}{\sqrt{k}} \qquad \alpha_k = D_{\mathcal{X}}/(M\sqrt{k})$$

Therefore, the choice of $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ yields average iterate convergence rate of $\mathcal{O}(\frac{1}{\epsilon^2})$

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).

[7] A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (Jan. 1, 2009). Publisher: Society for Industrial and Applied Mathematics, pp. 1574–1609. ISSN: 1052-6234. DOI: 10.1137/070704277. URL: https://epubs.siam.org/doi/abs/10.1137/070704277 (visited on 04/01/2020).

[8] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *arXiv:1606.04838 [cs, math, stat]* (Feb. 8, 2018). arXiv: 1606.04838. URL: http://arxiv.org/abs/1606.04838 (visited on 04/01/2020).