What is Momentum for Minimax Optimization?

Yuanhao Wang
IIIS, Tsinghua University

July 6, 2020

Abstract

By repeating the procedure of deriving momentum from conjugate gradient, one can derive from conjugate residual an algorithm similar to consensus optimization.

1 Warm-up: Deriving Momentum from Conjugate Gradient

Let us first see how we can derive momentum from CG. This can be found in Bach [2019] and Hardt [2018].

Consider a quadratic minimization problem $\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$, where $\mathbf{A}$ is positive definite with eigenvalues in the interval $[m, L]$. The solution to the problem is $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$, thus it is natural to consider using CG to solve it. Here, we won't use the actual update rule of CG, and would consider it as a Krylov subspace algorithm instead. Define the order-t Krylov subspace as

$$K_t(\mathbf{A}, \mathbf{b}) = \operatorname{span}\left\{\mathbf{b}, \mathbf{A}\mathbf{b}, \cdots, \mathbf{A}^{t-1}\mathbf{b}\right\}.$$

K_t can be thought as the space reachable via t gradient evaluations starting from the origin. The t-th iterate of CG is exactly

$$\mathbf{x}_t = \operatorname*{arg\,min}_{\mathbf{x}\in K_t(\mathbf{A},\mathbf{b})} \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} = \operatorname*{arg\,min}_{\mathbf{x}\in K_t(\mathbf{A},\mathbf{b})} \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\mathbf{A}(\mathbf{x}-\mathbf{x}^*).$$

Now, instead of considering the Krylov subspace, let us consider polynomials. $K_t(\mathbf{A}, \mathbf{b})$ is essentially the same as $\{p(\mathbf{A})\mathbf{b} | p \text{ is polynomial with degree} \le t-1\}$. Suppose that $\mathbf{x}_t = p(\mathbf{A})\mathbf{b}$, then $\mathbf{x}^* - \mathbf{x}_t = \mathbf{x}^* - p(\mathbf{A})\mathbf{b} = (\mathbf{I} - p(\mathbf{A})\mathbf{A})\mathbf{x}^*$. If p can be an arbitrary polynomial with degree $\le t-1$, then $q(x) = 1 - xp(x)$ can be an arbitrary polynomial in the set

$$\Pi_t := \{\text{Polynomials with degree} \le t \text{ such that } q(0) = 1\}.$$

In other words, in CG, $\mathbf{x}^* - \mathbf{x}_t = q_t(\mathbf{A})\mathbf{x}^*$, where

$$q_t \leftarrow \operatorname*{arg\,min}_{q\in\Pi_t} \left\{(\mathbf{x}^*)^T q_t(\mathbf{A})\mathbf{A}q_t(\mathbf{A})\mathbf{x}^*\right\}$$

Thus

$$\begin{aligned}(\mathbf{x}_t - \mathbf{x}^*)^T\mathbf{A}(\mathbf{x}_t - \mathbf{x}^*) &= \min_{q\in\Pi_t} \|q(\mathbf{A})\mathbf{x}^*\|_{\mathbf{A}} \le \min_{q\in\Pi_t} \|q(\mathbf{A})\|_2 \|\mathbf{x}^*\|_{\mathbf{H}} \\ &= \|\mathbf{x}^*\|_{\mathbf{H}} \cdot \min_{q\in\Pi_t} \max_{\lambda\in sp(\mathbf{A})} |q(\lambda)|.\end{aligned}$$

Because the spectrum of $\mathbf{A}$ is contained in $[m, L]$,

$$\min_{q\in\Pi_t} \max_{\lambda\in sp(\mathbf{A})} |q(\lambda)| \le \min_{q\in\Pi_t} \max_{\lambda\in[m,L]} |q(\lambda)|.$$

By polynomial approximation theory,

$$\min_{q\in\Pi_t}\max_{\lambda\in[m,L]}|q(\lambda)|\leq 2\left(\frac{\sqrt{\frac{L}{m}}-1}{\sqrt{\frac{L}{m}}+1}\right)^t,$$

which gives the well-known bound on CG's convergence. What we have done so far is upper bounding the performance of CG by the minimax polynomial. What if we use the minimax polynomial directly?

To answer this question, we'll have to look at the minimax polynomial on $[m,L]$, which is given by

$$q_t(z):=\frac{C_t(h(z))}{C_t(h(0))},\quad h(z):=\frac{2z-m-L}{L-m},$$

where $C_t(\cdot)$ is the t-th Chebyshev polynomial. q_t is essentially the t-th Chebyshev polynomial translated then rescaled. Chebyshev polynomials satisfies the recursion property:

$$C_{t+1}(x)=2zC_t(x)-C_{t-1}(x).$$

We can exploit this to get a recursion on q_t:

$$q_{t+1}(x)=\frac{2h(x)C_t(h(0))}{2h(0)C_t(h(0))-C_{t-1}(h(0))}\cdot q_t(x)-\frac{C_{t-1}(h(0))}{2h(0)C_t(h(0))-C_{t-1}(h(0))}\cdot q_{t-1}(x).$$

Assuming $L>>m$ and using asymptotic properties of Chebyshev polynomials, we get

$$q_{t+1}(x)\approx q_t(x)-\frac{2}{L}xq_t(x)+\left(1-4\sqrt{\frac{m}{L}}\right)(q_t(x)-q_{t-1}(x)).$$

This looks awfully familiar. Indeed, by plugging back into $\mathbf{x}^*-\mathbf{x}_t=q_t(\mathbf{A})\mathbf{x}^*$, this is exactly Polyak's momentum (heavy-ball method):

$$\mathbf{x}_{t+1}\approx\mathbf{x}_t-\frac{2}{L}\mathbf{A}(\mathbf{x}_t-\mathbf{x}^*)+\left(1-4\sqrt{\frac{m}{L}}\right)(\mathbf{x}_t-\mathbf{x}_{t-1}).$$

2 What is Momentum for Minimax Optimization?

Now, what happens if we repeat the same procedure for minimax optimization?

Consider a quadratic saddle point problem

$$\min_{\mathbf{x}}\max_{\mathbf{y}}f(\mathbf{x},\mathbf{y}):=\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{x}^T\mathbf{B}\mathbf{y}-\frac{1}{2}\mathbf{y}^T\mathbf{C}\mathbf{y}+\mathbf{u}^T\mathbf{x}+\mathbf{v}^T\mathbf{y}.$$

Let $\mathbf{H}:=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & -\mathbf{C}\end{bmatrix}$, $\mathbf{b}=-\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}$, and $\mathbf{z}=[\mathbf{x};\mathbf{y}]$. We assume that $\mathbf{A}\in\mathbb{R}^{n\times n}\succcurlyeq m_{\mathbf{x}}\mathbf{I}$, $\mathbf{C}\in\mathbb{R}^{m\times m}\succcurlyeq m_{\mathbf{y}}\mathbf{I}$, $\mathbf{B}\in\mathbb{R}^{n\times m}$, $\|\mathbf{H}\|_2\leq L$.

Fact 1. $\mathbf{H}:=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & -\mathbf{C}\end{bmatrix}$ *is indefinite, and the its eigenvalues fall in* $[-L,-m_{\mathbf{y}}]\cup[m_{\mathbf{x}},L]$.

The unique saddle point in this case is $\mathbf{z}^*=\mathbf{H}^{-1}\mathbf{b}$. However, one can no longer us CG to invert $\mathbf{H}$, since it is indefinite. An alternative is to use **Conjugate Residual** [Hestenes et al., 1952], which can be used to solve linear systems that are symmetric but indefinite. Conjugate Residual is also a Krylov subspace algorithm. Define

$$K_t(\mathbf{H},\mathbf{b})=\operatorname{span}\left\{\mathbf{b},\mathbf{H}\mathbf{b},\cdots,\mathbf{H}^{t-1}\mathbf{b}\right\},$$

then CR generates

$$\mathbf{z}_t \leftarrow \operatorname*{arg\,min}_{\mathbf{z}\in K_t(\mathbf{H},\mathbf{b})} \|\mathbf{H}\mathbf{z}-\mathbf{b}\|_2.$$

Converting to polynomials, this is equivalent to

$$q = \operatorname*{arg\,min}_{q\in\Pi_t} \|q(\mathbf{H})\mathbf{b}\|, \quad \mathbf{H}\mathbf{z}_t - \mathbf{b} = q(\mathbf{H})\mathbf{b}.$$

Here Π_t is the set of polynomials with degree at most t such that $q(0)=1$. Note that

$$\begin{aligned}\min_{q\in\Pi_t} \|q(\mathbf{H})\mathbf{b}\| &\le \min_{q\in\Pi_t} \|q(\mathbf{H})\|_2 \cdot \|\mathbf{b}\|_2 \\ &= \min_{q\in\Pi_t} \max_{\lambda\in sp(\mathbf{H})} |q(\lambda)| \cdot \|\mathbf{b}\|_2 \\ &\le \min_{q\in\Pi_t} \max_{\lambda\in[-L,-\mu_{\mathbf{y}}]\cup[\mu_{\mathbf{x}},L]} |q(\lambda)| \cdot \|\mathbf{b}\|_2.\end{aligned}$$

Therefore we can now upper bound the convergence of Conjugate Residual with the performance with the minimax polynomial on the union of two intervals. According to Greenbaum [1997, 3.13], the t-th order min-max polynomial for $[-L,-\mu_{\mathbf{y}}]\cup[\mu_{\mathbf{x}},L]$ is given by

$$q_t(z) := \frac{C_l(h(z))}{C_l(h(0))}, \quad h(z) := 1 + \frac{2(z+\mu_{\mathbf{y}})(z-\mu_{\mathbf{x}})}{-L^2+\mu_{\mathbf{x}}\mu_{\mathbf{y}}}, \quad l := \lfloor t/2 \rfloor$$

where$C_l(\cdot)$ is the l-th order Chebyshev polynomial. Now, let us try to recover a recurrence for $q_t(z)$.

$$\begin{aligned}q_{t+2}(z) &= \frac{2h(z)T_l(h(0))}{2h(0)C_l(h(0)) - C_{l-1}(h(0))} \cdot q_t(z) - \frac{C_{l-1}(h(0))}{2h(0)C_l(h(0)) - C_{l-1}(h(0))} \cdot q_{t-2}(z) \\ &\approx q_t(z) + \left(1 - \frac{\sqrt{m_{\mathbf{x}}m_{\mathbf{y}}}}{L}\right)(q_t(z) - q_{t-2}(z)) - \frac{m_{\mathbf{y}}-m_{\mathbf{x}}}{L^2} z q_t(z) - \frac{1}{L^2} z^2 q_t(z).\end{aligned}$$

Using $\mathbf{z}_t = \mathbf{z}^* + q(\mathbf{H})\mathbf{x}^*$, one derives the following update rule for $\mathbf{z}_t$:

$$\mathbf{z}_{t+1} \approx \mathbf{z}_t + \left(1 - \frac{\sqrt{m_{\mathbf{x}}m_{\mathbf{y}}}}{L}\right)(\mathbf{z}_t - \mathbf{z}_{t-1}) - \frac{m_{\mathbf{y}}-m_{\mathbf{x}}}{L^2}\mathbf{H}(\mathbf{z}_t - \mathbf{z}^*) - \frac{H^2}{L^2}\mathbf{H}^2(\mathbf{z}_t - \mathbf{z}^*). \tag{1}$$

Here, $\mathbf{H}^2(\mathbf{z}_t - \mathbf{z}^*) = \nabla(\frac{1}{2}\|\nabla f(\mathbf{x}_t,\mathbf{y}_t)\|^2)$, $\mathbf{H}(\mathbf{z}_t - \mathbf{z}^*) = \nabla f(\mathbf{x}_t,\mathbf{y}_t)$. So (1) can also be written as

$$\mathbf{z}_{t+1} \approx \mathbf{z}_t - \frac{1}{L^2}\nabla\left(\frac{1}{2}\|\nabla f(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) - \frac{m_{\mathbf{y}}-m_{\mathbf{x}}}{L^2}\nabla f(\mathbf{x}_t,\mathbf{y}_t) + \theta(\mathbf{z}_t - \mathbf{z}_{t-1}). \tag{2}$$

This is essentially gradient descent with momentum on a different loss

$$g(\mathbf{z}) := \frac{1}{2}\|\nabla f(\mathbf{x},\mathbf{y})\|^2 + (m_{\mathbf{y}} - m_{\mathbf{x}})f(\mathbf{x},\mathbf{y}),$$

This algorithm looks very similar to consensus optimization Mescheder et al. [2017], except that apart from doing gradient descent on the gradient norm, it does *gradient descent or ascent* on f, depending which of $m_{\mathbf{x}}$ and $m_{\mathbf{y}}$ is larger. In comparison, consensus optimization does gradient descent on the gradient norm, but does gradient descent-ascent on f.

This subtle difference may mean something. There is one very interesting fact about $g(\mathbf{z})$. In the quadratic case, $\frac{1}{2}\|\nabla f(\mathbf{x},\mathbf{y})\|^2$ is $\min\{m_{\mathbf{x}}, m_{\mathbf{y}}\}^2$-strongly convex, while $f(\mathbf{x},\mathbf{y})$ is neither convex nor concave. However, when they are added together, $g(\mathbf{z})$ becomes $m_{\mathbf{x}}m_{\mathbf{y}}$-strongly convex. When $m_{\mathbf{x}} \gg m_{\mathbf{y}}$ or $m_{\mathbf{x}} \ll m_{\mathbf{y}}$, the condition number of $g(\mathbf{z})$ is much smaller than that of $\frac{1}{2}\|\nabla f(\mathbf{x},\mathbf{y})\|^2$.

Fact 2. *$g(\mathbf{z})$ is $m_{\mathbf{x}}m_{\mathbf{y}}$-strongly convex and $2L^2$-smooth.*

This implies that by minimizing $g(\mathbf{z})$ instead of the gradient norm, one gets of rate of $O\left(\sqrt{\frac{L^2}{m_{\mathbf{x}}m_{\mathbf{y}}}}\ln\left(\frac{1}{\epsilon}\right)\right)$ instead of $O\left(\left(\frac{L}{m_{\mathbf{x}}}+\frac{L}{m_{\mathbf{y}}}\right)\ln\left(\frac{1}{\epsilon}\right)\right)$. (By the way, the lower bound for quadratic minimax optimization is $\Omega\left(\sqrt{\frac{L^2}{m_{\mathbf{x}}m_{\mathbf{y}}}}\ln\left(\frac{1}{\epsilon}\right)\right)$ [Zhang et al., 2019].)

Proof. The Hessian of $g(\mathbf{z})$ is $\mathbf{H}^2+(m_{\mathbf{y}}-m_{\mathbf{x}})\mathbf{H}$. Suppose λ is an eigenvalue of $\mathbf{H}$, then the corresponding eigenvalue of $\mathbf{H}^2+(m_{\mathbf{y}}-m_{\mathbf{x}})\mathbf{H}$ is $r(\lambda)=\lambda^2+(m_{\mathbf{y}}-m_{\mathbf{x}})\lambda$. Obviously

$$-m_{\mathbf{y}} < -\frac{m_{\mathbf{y}}-m_{\mathbf{x}}}{2} < m_{\mathbf{x}},$$

thus when $\lambda\in[-L,-m_{\mathbf{y}}]\cup[m_{\mathbf{x}},L]$, the minimum of $\lambda^2+(m_{\mathbf{y}}-m_{\mathbf{x}})\lambda$ is achieved when $\lambda=-m_{\mathbf{y}}$ or $\lambda=m_{\mathbf{x}}$. Hence

$$\lambda^2+(m_{\mathbf{y}}-m_{\mathbf{x}})\lambda \geq r(-m_{\mathbf{y}})=r(m_{\mathbf{x}})=m_{\mathbf{x}}m_{\mathbf{y}}.$$

Meanwhile,

$$|\lambda^2+(m_{\mathbf{y}}-m_{\mathbf{x}})\lambda| \leq L^2+|m_{\mathbf{y}}-m_{\mathbf{x}}|L \leq 2L^2.$$

□

References

Francis Bach. Polynomial magic I : Chebyshev polynomials. https://francisbach.com/chebyshev-polynomials/, Nov 2019.

Anne Greenbaum. *Iterative methods for solving linear systems*, volume 17. SIAM, 1997.

Moritz Hardt. Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation . https://ee227c.github.io/notes/ee227c-lecture06.pdf, Oct 2018.

Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.