Blog Post #3 March 17, 2026

Online Learning to Precondition I. Space dilation methods for linear systems

Wenzhi Gao 

gwz@stanford.edu ICME, Stanford University

Abstract

This post begins a series on what I call Online Learning to Precondition (OL2P). For a mathematical optimization problem, the core idea is to

improve the optimization landscape by learning from the behavior of the algorithm itself.

This principle is implicit in several existing optimization algorithms and can lead to powerful results. However, a systematic study of this idea seems to be lacking. In this series, I will explain the principle and, more importantly, give concrete instantiations, drawn both from the literature and from my own work. As the first example, I present a "new" algorithm for solving linear systems. The method has superlinear convergence and resembles a quasi-Newton method, but it is derived entirely from a learning principle.

Updates.

- 03/17/2026. Fix typos.

1 Introduction and background

Consider the convex quadratic minimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) := \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle,$$

where $A \in \mathbb{S}^n_{++}$ is positive definite and contains eigenvalues $\lambda_1 \geq \lambda_2, \ldots, \lambda_n$. Its optimality condition $0 = \nabla f(x) = Ax - b$ corresponds to solving a linear system. For simplicity, I assume that $f(x^\star)$ is known. It can be achieved, for example, with a least-squares formulation $f(x) = \frac{1}{2}\|Ax - b\|^2$.

1.1 Optimization with steepest descent

A standard algorithm for solving this problem is steepest descent

$$\begin{aligned} \alpha_k &= \arg\min_{\alpha} f(x^k - \alpha \nabla f(x^k)) = \frac{\|\nabla f(x^k)\|^2}{\langle \nabla f(x^k), A\nabla f(x^k)\rangle} \\ x^{k+1} &= x^k - \alpha_k \nabla f(x^k), \end{aligned}$$

where α_k minimizes f over the line $\{x : x = x^k + \alpha \nabla f(x^k), \alpha \in \mathbb{R}\}$. From now on, let's focus on a single step (from x to x^+) and drop the iteration index. Since f is a quadratic, we can explicitly write down the contraction ratio, denoted by r, as

$$r := \frac{f(x^+) - f(x^\star)}{f(x) - f(x^\star)} = 1 - \frac{\|\nabla f(x)\|^4}{\|\nabla f(x)\|_A^2 \|\nabla f(x)\|_{A^{-1}}^2} = 1 - \frac{1}{\left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, A\frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, A^{-1}\frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle}$$

which, by Kantorovich's inequality $\frac{\|a\|^2}{\|a\|_A \|a\|_{A^{-1}}} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}$, gives $r \leq 1 - \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2} = \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2$. When $\lambda_1 \gg \lambda_n$, the problem is called ill-conditioned, as it is possible that $r \to 1$.

More detailedly, this issue of bad contraction ratio can happen if at least one of $\left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, A\frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle$ or $\left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, A^{-1}\frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle$ is large. Denote $g := \frac{\nabla f(x)}{\|\nabla f(x)\|}$. Then

The contraction ratio r is bad if at least one of $\langle g, Ag\rangle$ or $\langle g, A^{-1}g\rangle$ is large.

A standard technique for dealing with ill-conditioned problems is preconditioning.

1.2 Preconditioning

There are many ways to write preconditioning. I'll use a very much simplified version. Let P be a positive definite matrix. Then solving the original problem is the same as solving

$$\min_x \quad \tfrac{1}{2}\langle x, PAPx\rangle - \langle Pb, x\rangle$$

up to a change of variable $x \leftarrow Px$. This new function has Hessian PAP, and the choice of P should ideally make PAP look like identity, say, $\operatorname{dist}(PAP, I) \to 0$.

The question is how to choose P. Let's start by choosing a distance function to measure the distance between a positive definite matrix A and identity I. There are many matrix functions available for measuring distance (e.g., Frobenius norm $\|P-Q\|_F$). Here I use the log det divergence (although strictly speaking it is a divergence rather than a distance):

$$V_d(P, Q) = \langle P, Q^{-1}\rangle - \log\det PQ^{-1} - n,$$

which is the Bregman divergence $V_d(P,Q) = d(P) - d(Q) - \langle \nabla d(Q), P-Q\rangle$ induced by $d(X) = -\log\det X$. There are many technical details regarding log det divergence, but it will just be used as a distance measure here: we want to find P such that

$$\operatorname{dist}(PAP, I) = V_d(PAP, I) = \operatorname{tr}(PAP) - \log\det PAP - n$$

is close to 0. Let's also assume P has some parametrized form $P = I + \sigma vv^\top$ for some unit vector $\|v\| = 1$. This form of P is known as space-dilation operator in the optimization literature, since when it operates on a vector, it simply stretches (squeezes) the space in direction v:

$$Pz = (I + \sigma vv^\top)z = z + \sigma\langle v, z\rangle v.$$

Given $P = I + \sigma vv^\top$, we can write $A^+(v,\sigma) := PAP = (I+\sigma vv^\top)A(I+\sigma vv^\top)$. Using the log det divergence, the way $V_d(PAP, I)$ changes w.r.t. $V_d(A, I)$ can be explicitly computed:

$$\begin{aligned}
\operatorname{tr}(PAP) &= \operatorname{tr}((I+\sigma vv^\top)(A + \sigma Avv^\top)) \\
&= \operatorname{tr}(A + \sigma Avv^\top + \sigma vv^\top A + \sigma^2 vv^\top Avv^\top) \\
&= \operatorname{tr}(A) + 2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle \\
\det PAP &= \det(I+\sigma vv^\top)A(I+\sigma vv^\top) \\
&= (1+\sigma)^2 \det A.
\end{aligned}$$

Hence we can explicitly write

$$\begin{aligned}
V_d(A^+(v,\sigma), I) &= V_d(PAP, I) \\
&= \operatorname{tr}(A) + 2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle - \log[(1+\sigma)^2\det A] - n \\
&= \operatorname{tr}(A) - \log\det A - n + 2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle - 2\log(1+\sigma) \\
&= V_d(A, I) + \underbrace{2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle - 2\log(1+\sigma)}_{\gamma(\sigma)},
\end{aligned}$$

where $\gamma(\sigma)$ is a scalar function in σ given v. Since our goal is to reduce $V_d(PAP, I)$, we should minimize $\gamma(\sigma)$, whose optimal value is $\sigma^\star = \sqrt{\frac{1}{\langle v, Av\rangle}} - 1$ and

$$V_d(A^+(v,\sigma^\star), I) = V_d(A, I) + \underbrace{\log\left(\langle v, Av\rangle\right) - \langle v, Av\rangle + 1}_{\eta(\langle v, Av\rangle)},$$

where $\eta(x) = \log x - x + 1 \le 0$ is again a scalar function. This function is always non-positive, with maximizer attained at $x = 1$. Hence if we can find v such that $\langle v, Av\rangle$ is bounded away from 1 (e.g. $|\langle v, Av\rangle - 1| \ge c_1 > 0$), then we can ensure that

$$V_d(A^+(v,\sigma^\star), I) \le V_d(A, I) - c_2$$

for some other constant $c_2 > 0$. In other words, direction v helps to reduce the distance. Since $\|v\| = 1$, having $|\langle v, Av\rangle - 1| > c_1$ requires v to have either large or small Rayleigh quotient. The only question is where we can find such v? Let's rephrase the problem.

The distance can improve if we find v such that $\langle v, Av\rangle$ is bounded away from 1.

1.3 Learning to precondition

Let's put together the observations from the previous sections. After one steepest descent step, we have

$$r = 1 - \frac{1}{\left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, A\frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|}, A^{-1}\frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle} = 1 - \frac{1}{\langle g, Ag\rangle\langle g, A^{-1}g\rangle}.$$

- **Case 1**. If r not too large (say $r \le 0.99$), then $\langle g, Ag\rangle$ and $\langle g, A^{-1}g\rangle$ are both small.
- **Case 2**. If $r \to 1$ (say $r > 0.99$), then $\frac{1}{\langle g, Ag\rangle\langle g, A^{-1}g\rangle} \to 0$ and at least one of them is large.

  But the distance can improve if we find v such that $\langle v, Av\rangle$ is bounded away from 1.

Clearly, there is no need to worry when the contraction is already good. The interesting case is **Case 2** $r \to 1$, where we know at least one of $\langle g, Ag\rangle$ and $\langle g, A^{-1}g\rangle = \langle (A^{-1}g), A(A^{-1}g)\rangle$ is large.

- If $\langle g, Ag\rangle \gg 1$, then we can let $v \leftarrow g$ and ensure $V_d(A^+(v,\sigma^\star), I) \le V_d(A, I) - c_2$.
- If $\langle g, A^{-1}g\rangle \gg 1$, then we can let $v \leftarrow A^{-1}g$ and again ensure $V_d(A^+(v,\sigma^\star), I) \le V_d(A, I) - c_2$.

Therefore, every time **Case 2** happens, we always ensure that

$$V_d(A^+(v,\sigma^\star), I) \le V_d(A, I) - c_2.$$

Suppose we replace $A \leftarrow A^+(v,\sigma^\star), b \leftarrow Pb$ whenever **Case 2** happens and scale x by $x \leftarrow P^{-1}x$, then the function value remains unchanged:

$$\tfrac{1}{2}\langle (P^{-1}x), PAP(P^{-1}x)\rangle - \langle Pb, P^{-1}x\rangle = \tfrac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle.$$

Finally, the nonnegativity of V_d implies **Case 2** at most happen $\left\lceil \frac{V_d(A,I)}{c_2} \right\rceil$ times. the final complexity of steepest descent becomes

$$\mathcal{O}\left(V_d(A, I) + \log\left(\tfrac{1}{\varepsilon}\right)\right).$$

Algorithm 1. Space dilation for solving linear system

input A, b, x and $f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$
for $k = 1, \ldots, K$
  compute $\alpha = \frac{\|\nabla f(x)\|^2}{\langle \nabla f(x), A\nabla f(x)\rangle}$ and $r = \frac{f(x - \alpha\nabla f(x)) - f(x^\star)}{f(x) - f(x^\star)}, g = \frac{\nabla f(x)}{\|\nabla f(x)\|}$
  if $r > 0.99$
    if $\langle g, Ag\rangle$ is large
      $v = g$
    else ($\langle g, A^{-1}g\rangle$ must be large)
      $v = A^{-1}g$
    end if
    compute $\sigma^\star = \sqrt{\frac{1}{\langle v, Av\rangle}} - 1$
    let $P = I + \sigma^\star vv^\top$
    let $A \leftarrow PAP, b \leftarrow Pb, x \leftarrow P^{-1}x$
  end if
end for

Theorem 1. ***Algorithm 1*** *finds an ε-optimal solution in $\mathcal{O}\left(V_d(A, I) + \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations.*

The algorithm above is clearly not implementable due to the need for computing $A^{-1}g$, but it is sufficient as a prototypical method for communicating intuitions. In the next section, I'll remove the need for evaluating $A^{-1}g$ by using symmetrized log det divergence.

Theorem 2. *A variant of **Algorithm 1** (in the next section) finds an ε-optimal solution in $\mathcal{O}\left(V_d(A, I) + V_d(I, A) + \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations. Moreover, the algorithm has superlinear convergence*

$$\frac{f(x^{K+1}) - f(x^\star)}{f(x^1) - f(x^\star)} \leq \left(\frac{C}{K}\right)^K$$

for some $C > 0$.

As a final remark, the idea mentioned in this post was pioneered by [MON04] and [DH07]. Both papers' authors notice that when an algorithm performs poorly (gives bad contraction), it simultaneously provides information that helps improve problem conditioning (reduce distance). The approach in this post is inspired by [DH07], and our log det-based analysis partially resolves some shortcomings mentioned at the end of [DH07].

2 Space dilation methods for linear systems

Algorithm 1 in the previous section is not implementable due to the need of computing $A^{-1}g$. This section fixes this issue. When $A^{-1}g$ is not available, we need to deal with the case $\langle g, A^{-1}g\rangle \gg 1$ and $\langle g, Ag\rangle \to 1$.

The key observation is simple:

$$\text{when } \langle g, Ag\rangle \gg 1, \text{ the vector } g \text{ can help improve conditioning of } A$$

which, symmetrically, should give

$$\text{when } \langle g, A^{-1}g\rangle \gg 1, \text{ the vector } g \text{ can help improve conditioning of } A^{-1}.$$

So let's define

$$\text{dist}(PAP, I) = \underbrace{V_d(PAP, I)}_{PAP \text{ to } I} + \underbrace{V_d((PAP)^{-1}, I)}_{(PAP)^{-1} \text{ to } I},$$

which, by the property of log det divergence $V_d(P,Q) = V_d(Q^{-1}, P^{-1})$, gives

$$\mathrm{dist}(PAP, I) = V_d(PAP, I) + V_d(I, PAP).$$

It is exactly the symmetrized version of the divergence. Now we can repeat the previous derivation with $P = I + \sigma vv^\top$ and $P^{-1} = I - \frac{\sigma}{1+\sigma} vv^\top$:

$$\begin{aligned} V_d(PAP, I) &= \mathrm{tr}(A) - \log\det A - n + 2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle - 2\log(1+\sigma) \\ &= V_d(A, I) + 2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle - 2\log(1+\sigma). \\ V_d(I, PAP) &= V_d(I, A) - \tfrac{2\sigma}{1+\sigma}\langle v, A^{-1}v\rangle + \tfrac{\sigma^2}{(1+\sigma)^2}\langle v, A^{-1}v\rangle + 2\log(1+\sigma), \end{aligned}$$

and we have

$$\begin{aligned} & V_d(PAP, I) + V_d(I, PAP) \\ \le\ & V_d(A, I) + V_d(I, A) + 2\sigma\langle v, Av\rangle + \sigma^2\langle v, Av\rangle - 2\log(1+\sigma) \\ & -\tfrac{2\sigma}{1+\sigma}\langle v, A^{-1}v\rangle + \tfrac{\sigma^2}{(1+\sigma)^2}\langle v, A^{-1}v\rangle + 2\log(1+\sigma) \\ =\ & V_d(A, I) + V_d(I, A) + \underbrace{(\sigma^2 + 2\sigma)\langle v, Av\rangle - \tfrac{\sigma^2+2\sigma}{(1+\sigma)^2}\langle v, A^{-1}v\rangle}_{\gamma(\sigma)}. \end{aligned}$$

Minimizing over σ, we have $\sigma^\star = -1 + \left(\frac{\langle v, A^{-1}v\rangle}{\langle v, Av\rangle}\right)^{1/4}$ and that

$$V_d(A^+(v,\sigma^\star), I) + V_d(I, A^+(v,\sigma^\star)) \le V_d(A, I) + V_d(I, A) - \left(\sqrt{\langle v, A^{-1}v\rangle} - \sqrt{\langle v, Av\rangle}\right)^2.$$

The RHS is now the Hellinger distance between two scalars $\langle v, A^{-1}v\rangle$ and $\langle v, Av\rangle$. To ensure that

$$\left|\sqrt{\langle v, A^{-1}v\rangle} - \sqrt{\langle v, Av\rangle}\right| \ge c_1,$$

we need $\max\{\langle v, A^{-1}v\rangle, \langle v, Av\rangle\} - \min\{\langle v, A^{-1}v\rangle, \langle v, Av\rangle\} \ge c_1$, i.e., v achieves sufficiently different Rayleigh quotients on A and A^{-1}. Can we find such v when $r = 1 - \frac{1}{\langle g, Ag\rangle\langle g, A^{-1}g\rangle} \to 1$? Let's do a concrete case analysis.

Case 1. If $r \le \frac{15}{16}$, then we are fine.

Case 2. If $r > \frac{15}{16}$, then $\frac{1}{\langle g, A^{-1}g\rangle\langle g, Ag\rangle} \le \frac{1}{16}$ and $\max\{\langle g, Ag\rangle, \langle g, A^{-1}g\rangle\} \ge 4$.

- **Case 2.1**. If $\min\{\langle g, Ag\rangle, \langle g, A^{-1}g\rangle\} \le 2$. Then $\left|\sqrt{\langle g, A^{-1}g\rangle} - \sqrt{\langle g, Ag\rangle}\right| \ge 2$. We can just take $v = g$.
- **Case 2.2**. If $\min\{\langle g, Ag\rangle, \langle g, A^{-1}g\rangle\} \ge 2$. This case is slightly more subtle. It is possible both $\langle g, Ag\rangle$ and $\langle g, A^{-1}g\rangle$ are equally large. Interestingly, we can show that running a power iteration on top of g would generate a new direction that meets our need: in particular, let $z = \frac{Ag}{\|Ag\|}$. We have, by spectral Cauchy-Schwarz, that

$$\begin{aligned} \langle z, Az\rangle = \tfrac{\langle g, A^3 g\rangle}{\|Ag\|^2} &\ge \tfrac{\|Ag\|^2\langle g, Ag\rangle}{\|Ag\|^2} = \langle g, Ag\rangle \ge 2 \\ \langle z, A^{-1}z\rangle = \tfrac{\langle g, Ag\rangle}{\|Ag\|^2} &\le \tfrac{\langle g, Ag\rangle}{\langle g, Ag\rangle^2} \le \tfrac{1}{2} \end{aligned}$$

  and taking $v = z$ gives $\left|\sqrt{\langle z, A^{-1}z\rangle} - \sqrt{\langle z, Az\rangle}\right| \ge \sqrt{2} - \frac{\sqrt{2}}{2} \ge \frac{1}{4}$.

Now we have **Algorithm 2**.

Algorithm 2. Another space dilation algorithm

input A, b, x and $f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$
for $k = 1, \ldots, K$
  compute $\alpha = \frac{\|\nabla f(x)\|^2}{\langle \nabla f(x), A\nabla f(x)\rangle}$ and $r = \frac{f(x-\alpha\nabla f(x)) - f(x^\star)}{f(x) - f(x^\star)}, g = \frac{\nabla f(x)}{\|\nabla f(x)\|}$
  if $r > \frac{15}{16}$
    if Case 2.1 happens
      $v = g$
    else
      $v = \frac{Ag}{\|Ag\|}$
    end if
    compute $\sigma^\star = \left(\frac{\langle v, A^{-1}v\rangle}{\langle v, Av\rangle}\right)^{1/4} - 1$
    let $P = I + \sigma^\star vv^\top$
    let $A \leftarrow PAP, b \leftarrow Pb, x \leftarrow P^{-1}x$
  end if
end for

As a last minor point, **Algorithm 2** requires computing $\langle v, A^{-1}v\rangle$ in $\sigma^\star$. When $v = \frac{Ag}{\|Ag\|}$, this is not a problem, since $\left\langle \frac{Ag}{\|Ag\|}, A^{-1}\frac{Ag}{\|Ag\|}\right\rangle = \frac{\langle g, Ag\rangle}{\|Ag\|^2}$. The only issue happens when $v = g$, where

$$\frac{1}{2}\langle g, A^{-1}g\rangle = \frac{1}{2}\frac{\langle \nabla f(x), A^{-1}\nabla f(x)\rangle}{\|\nabla f(x)\|^2} = \frac{f(x) - f(x^\star)}{\|\nabla f(x)\|^2}$$

holds for quadratic functions. Although no $A^{-1}g$ is required, it does depend on the optimality gap.

2.1 Potential reduction

To prove the promised **Theorem 2**, I'll introduce a tool for analyzing algorithms with this "learning to optimize" intuition. At each step of the algorithm, we

- either make progress (contraction $r \to 0$)
- or improve the landscape (V_d shrinks).

Thus, even when the contraction is poor, each step still decreases a suitable potential.

For the specific algorithm we consider, let's define the potential function

$$\varphi(A, x) \coloneqq \underbrace{\log(f(x) - f(x^\star))}_{\text{Optimization}} + \underbrace{[V_d(A, I) + V_d(I, A)]}_{\text{Learning}}.$$

The algorithm is slightly different, with a different threshold for choosing between two cases.

Algorithm 3. Space dilation with superlinear convergence rate

input A, b, x and $f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$
for $k = 1, \ldots, K$
  compute $\alpha = \frac{\|\nabla f(x)\|^2}{\langle \nabla f(x), A\nabla f(x)\rangle}$ and $r = \frac{f(x-\alpha\nabla f(x)) - f(x^\star)}{f(x) - f(x^\star)}, g = \frac{\nabla f(x)}{\|\nabla f(x)\|}$
  if $\min\{\langle g, Ag\rangle, \langle g, A^{-1}g\rangle\} \le \sqrt{\frac{2-r}{2(1-r)}}$
    $v = g$
  else
    $v = \frac{Ag}{\|Ag\|}$
  end if
  compute $\sigma^\star = \left(\frac{\langle v, A^{-1}v\rangle}{\langle v, Av\rangle}\right)^{1/4} - 1$
  let $P = I + \sigma^\star vv^\top$
  let $A \leftarrow PAP, b \leftarrow Pb, x \leftarrow P^{-1}x$
end for

Theorem 3. ***Algorithm 3*** *satisfies* $\varphi(A^+,x^+) \le \varphi(A,x) - \frac{1}{2}\log\frac{\mathrm{e}}{2}$. *Moreover, superlinear convergence holds:*

$$\frac{f(x^{K+1})-f(x^\star)}{f(x^1)-f(x^\star)} \le \left(\sqrt{\frac{4[V_d(A,I)+V_d(I,A)]}{K}}\right)^K.$$

Proof. Since $r = 1 - \frac{1}{\langle g,Ag\rangle\langle g,A^{-1}g\rangle}$, we have $\langle g,Ag\rangle\langle g,A^{-1}g\rangle = \frac{1}{1-r}$ and $\max\{\langle g,Ag\rangle,\langle g,A^{-1}g\rangle\} \ge \frac{1}{\sqrt{1-r}}$.

Case 1. If $\min\{\langle g,Ag\rangle,\langle g,A^{-1}g\rangle\} \le \alpha := \sqrt{\frac{2-r}{2(1-r)}}$, $v = g$ gives $\max\{\langle g,Ag\rangle,\langle g,A^{-1}g\rangle\} \ge \max\left\{\frac{1}{1-r}\alpha^{-1}, \frac{1}{\sqrt{1-r}}\right\}$,

$$\begin{aligned}
\left(\sqrt{\langle v,A^{-1}v\rangle} - \sqrt{\langle v,Av\rangle}\right)^2 &\ge \max\left\{\max\left\{\frac{1}{1-r}\alpha^{-1}, \frac{1}{\sqrt{1-r}}\right\} - \alpha, 0\right\}^2 \\
&= \max\left\{\max\left\{\sqrt{\frac{2}{(2-r)(1-r)}}, \frac{1}{\sqrt{1-r}}\right\} - \sqrt{\frac{2-r}{2(1-r)}}, 0\right\}^2 \\
&= \max\left\{\max\left\{\sqrt{\frac{2}{(2-r)}}\frac{1}{\sqrt{1-r}}, \frac{1}{\sqrt{1-r}}\right\} - \sqrt{\frac{2-r}{2(1-r)}}, 0\right\}^2 \\
&= \max\left\{\sqrt{\frac{2}{2-r}}\frac{1}{\sqrt{1-r}} - \sqrt{\frac{2-r}{2}}\frac{1}{\sqrt{1-r}}, 0\right\}^2 \\
&= \frac{r^2}{2(r-2)(r-1)}
\end{aligned}$$

Case 2. If $\min\{\langle g,Ag\rangle,\langle g,A^{-1}g\rangle\} \ge \alpha$, $v = \frac{Ag}{\|Ag\|}$ gives $\langle v,Av\rangle \ge \alpha$, $\langle v,A^{-1}v\rangle \le \frac{1}{\alpha}$ and

$$\begin{aligned}
\left(\sqrt{\langle v,A^{-1}v\rangle} - \sqrt{\langle v,Av\rangle}\right)^2 &\ge \max\left\{\alpha - \frac{1}{\alpha}, 0\right\}^2 \\
&= \max\left\{\sqrt{\frac{2-r}{2-2r}} - \sqrt{\frac{2-2r}{2-r}}, 0\right\}^2 \\
&= \left(\sqrt{\frac{2-r}{2-2r}} - \sqrt{\frac{2-2r}{2-r}}\right)^2 = \frac{r^2}{2(r-2)(r-1)}.
\end{aligned}$$

Hence $\left(\sqrt{\langle v,A^{-1}v\rangle} - \sqrt{\langle v,Av\rangle}\right)^2 \ge \frac{r^2}{2(2-r)(1-r)} \ge \frac{r^2}{4}$ in both cases. In view of the potential function, we have

$$\begin{aligned}
\log\left(f(x^+)-f(x^\star)\right) &= \log\left(f(x)-f(x^\star)\right) + \log r \\
V_d(PAP,I) + V_d(I,PAP) &\le V_d(A,I) + V_d(I,A) - \frac{r^2}{4}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\varphi(A^+,x^+) &\le \varphi(A,x) + \log r - \frac{r^2}{4} \\
&\le \varphi(A,x) + \max_{r\in[0,1]} \log r - \frac{r^2}{4} \\
&= \varphi(A,x) - \frac{1}{2}\log\frac{\mathrm{e}}{2}
\end{aligned}$$

Finally, to show superlinear convergence, we denote $r_k = \frac{f(x^{k+1})-f(x^\star)}{f(x^k)-f(x^\star)}$, and the relation

$$V_d(PAP,I) + V_d(I,PAP) \le V_d(A,I) + V_d(I,A) - \frac{r^2}{4}$$

implies $\frac{1}{K}\sum_{k=1}^K r_k^2 \le 4[V_d(A,I)+V_d(I,A)]$, and with quadratic mean inequality, we have

$$\frac{1}{K}\sum_{k=1}^K r_k \le \sqrt{\frac{1}{K}\sum_{k=1}^K r_k^2} \le \sqrt{\frac{4[V_d(A,I)+V_d(I,A)]}{K}},$$

which, by AM-GM inequality, implies

$$\frac{f(x^{K+1})-f(x^\star)}{f(x^1)-f(x^\star)} \le \left(\sqrt{\frac{4[V_d(A,I)+V_d(I,A)]}{K}}\right)^K$$

and this completes the proof. □

References

[DH07] John Dunagan and Nicholas JA Harvey. Iteratively constructing preconditioners via the conjugate gradient method. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 207–216. 2007.

[MON04] Renato DC Monteiro, Jerome W O'Neal, and Arkadi Nemirovski. A new conjugate gradient algorithm incorporating adaptive ellipsoid preconditioning. *Report, School of ISyE, Georgia Tech, USA*, 2004.