# Notes and errata to "A proof of the exact convergence rate of gradient descent"

Jungbin Kim

March 10, 2026

## 1 More than a dry theorem

The paper [6] suggests that finding the "right" level of generality and discovering hidden structures are important parts of proving a theorem. The level of abstraction in [6] is very elementary. It is accessible to anyone with a basic knowledge of calculus and linear algebra.

The conjecture in [2] was formulated only for the case $\mu = 0$. Later, this conjecture was generalized to the case $\mu \geq 0$ in [10]. One might wonder if generalizing an unproven conjecture actually helps prove for the special case. When the case $\mu = 0$ is still unresolved, attempting the case $\mu \geq 0$ may appear even more inaccessible. But after reading [6, §4], one will see that the generalized conjecture (more precisely, its "mirror") is actually easier to handle, at least when following the proof I have presented. The reasoning of fixing $\gamma$ and varying $\mu$ and $L$ reveals a hidden analytic structure that is nearly impossible to see when considering only the special case $\mu = 0$.

A more fundamental structure, which is algebraic, can be found in [6, §3]. It suggests that in order to truly understand the anti-transpose relationship between optimization algorithms, one should view an algorithm as the operator $\tilde{\mathbf{H}}^{-1}$ rather than merely as a collection of stepsizes $\mathbf{H}_{i,j}$. By adopting this view, a proof of a general theorem was obtained. It is a new proof, not an extension of the previously known proof.

## 2 Heuristic evidence for considering the basis $\{x_{k+1}^+ - x_k^+\}$

To summarize the overall line of reasoning in [6, §3], PEPs are formulated using the basis $\{x_{k+1}^+ - x_k^+\}$ while involving the operator $\tilde{\mathbf{H}}^{-1}$ in the matrix that is shown to be positive semidefinite. Then, the equivalence between the two PEP analyses is established by using the identity $\langle \tilde{\mathbf{H}}^{-1}x, \mathrm{T}y\rangle = \langle \mathrm{T}x, \tilde{\mathbf{H}}^{-\mathrm{A}}y\rangle$.

Among these steps, the real breakthrough is recognizing the operator $\tilde{\mathbf{H}}^{-1}$. Simply changing the basis from $\{\nabla f(x_k)\}$ to $\{x_{k+1}^+ - x_k^+\}$ is a superficial step. Without a fundamental change in the proof structure, this will merely rewrite the same operators in new coordinates. However, once one believes that $\{x_{k+1}^+ - x_k^+\}$ is the "right" basis for PEP analyses and that the previously known proof might not be the "right" one, the most natural structure one might imagine as the underlying essence of the equivalence between the two PEP analyses is an identity like $\langle \mathrm{H}^{-1}x, \mathrm{T}y\rangle = \langle \mathrm{T}x, \mathrm{H}^{-\mathrm{A}}y\rangle$ (of course, one should check $(\mathrm{H}^{-1})^{\mathrm{A}} = (\mathrm{H}^{\mathrm{A}})^{-1}$).

There were three pieces of evidence that suggested $\{x_{k+1}^+ - x_k^+\}$ as the "right" basis for PEP analyses. The first one is the most direct. The second one came from findings by other researchers. The last one arose from my previous work. Other relevant ideas may exist in the literature that I might have missed. In particular, I have very little awareness of the literature appearing after 2024. I apologize to researchers whose contributions are not mentioned here and ask for their kind understanding.

## 2.1 Evidence 1: Extension to the strongly convex setting

Consider the following two continuous-time models:

$$\dot{X}(t) = -\int_0^t H^F(t,\tau)\nabla f(X(\tau))\,d\tau, \qquad \dot{Y}(t) = -\int_0^t H^G(t,\tau)\nabla f(Y(\tau))\,d\tau,$$

where the first one aims to minimize $(f(X(T)) - f(x_*))/\|x_0 - x_*\|^2$ and the second one aims to minimize $\|\nabla f(Y(T))\|^2/(f(y_0) - f(y_*))$.

Suppose that the convergence rate of the first model is proven by taking the weighted integral of

$$\begin{aligned}
0 &= \frac{d}{dt}\{f(X(t)) - f(x_*)\} - \left\langle \nabla f(X(t)), \dot{X}(t)\right\rangle && \text{with weight } \lambda^F(t),\\
0 &\geq f(X(t)) - f(x_*) - \langle \nabla f(X(t)), X(t) - x_*\rangle + \boldsymbol{\frac{\mu}{2}\|X(t) - x_*\|^2} && \text{with weight } \dot{\lambda}^F(t),
\end{aligned}$$

and that the convergence rate of the second model is proven by taking the weighted integral of

$$\begin{aligned}
0 &= \frac{d}{dt}\{f(Y(t)) - f(Y(T))\} - \left\langle \nabla f(Y(t)), \dot{Y}(t)\right\rangle && \text{with weight } \lambda^G(t),\\
0 &\geq f(Y(t)) - f(Y(T)) - \langle \nabla f(Y(t)), Y(t) - Y(T)\rangle + \boldsymbol{\frac{\mu}{2}\|Y(t) - Y(T)\|^2} && \text{with weight } \dot{\lambda}^G(t).
\end{aligned}$$

The previously known proof of the equivalence between the two convergence analyses (with $\lambda^F(t) = 1/\lambda^G(T-t)$) was done by formulating continuous-time PEPs in the bases $\{\nabla f(X(t))\}$ and $\{\nabla f(Y(t))\}$, and then identifying $\nabla f(X(T-t)) = \lambda^G(t)\nabla f(Y(t)) - \int_0^t \dot{\lambda}^G(\tau)\nabla f(Y(\tau))\,d\tau$ to show that the two PEP analyses are identical (see [4, Thm. 2] or [7, Thm. 3]).

But that proof seems to work nicely only when $\mu = 0$. If the known proof is not the "right" one, then in the "right" proof, one might hope that the bolded terms correspond in the sense that

$$\int_0^T \dot{\lambda}^F(t)\frac{\mu}{2}\|X(t) - x_*\|^2\,dt = \int_0^T \dot{\lambda}^G(t)\frac{\mu}{2}\|Y(t) - Y(T)\|^2\,dt.$$

Thus, the most natural identification (or change-of-basis formula) is (note that $\dot{\lambda}^G(t) = \dot{\lambda}^F(T-t)/\lambda^F(T-t)^2$)

$$X(t) - x_* = \frac{Y(T-t) - Y(T)}{\lambda^F(t)},$$

suggesting that the two PEPs should be formulated in the $\{X(t)\}$ and $\{Y(t)\}$ bases (or their variants) rather than the $\{\nabla f(X(t))\}$ and $\{\nabla f(Y(t))\}$ bases.

## 2.2 Evidence 2: Correspondence between Lyapunov analyses

I will start with a concrete example to motivate a general result. The order of presentation here may differ from the actual chronological development of [1].

**Example 1** ([8]). *Fix a constant $r \geq 3$. Consider the ODE models*

$$\ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X) = 0, \qquad \ddot{Y} + \frac{r}{T-t}\dot{Y} + \nabla f(Y) = 0.$$

*Consider the energy functions*

$$\begin{aligned}
\mathcal{E}^F(t) &= t^2(f(X(t)) - f(x^*)) + \frac{1}{2}\left\|t\dot{X}(t) + 2(X(t) - x_*)\right\|^2 + (r-3)\left\|X(t) - x_*\right\|^2, \\
\mathcal{E}^G(t) &= \frac{1}{(T-t)^2}\left(f(Y(t)) - f(Y(T))\right) + \frac{1}{2(T-t)^4}\left\|(T-t)\dot{Y}(t) + 2(Y(t) - Y(T))\right\|^2 \\
&\quad - \frac{r-1}{(T-t)^4}\left\|Y(t) - X(T)\right\|^2.
\end{aligned}$$

*By using convexity and performing a calculation, one will obtain the following inequalities:*

$$\begin{aligned}
\dot{\mathcal{E}}^F(t) &\leq -(r-3)t\left\|\dot{X}(t)\right\|^2, \\
\dot{\mathcal{E}}^G(T-t) &\leq -\frac{r-3}{t^5}\left\|t\dot{Y}(T-t) + 2(Y(T-t) - Y(T))\right\|^2.
\end{aligned} \tag{1}$$

One can observe that the right-hand sides of (1) are the same under the identification $X(t) - x_* = (Y(T-t) - Y(T))/t^2$. The following theorem generalizes this observation.

**Theorem 1** ([1]). *Consider the following two ODE models:*

$$\ddot{X} + a(t)\dot{X} + \nabla f(X) = 0, \qquad \ddot{Y} + a(T-t)\dot{Y} + \nabla f(Y) = 0$$

*Then, in some sense, there is a correspondence between two specific Lyapunov analyses.*

*Proof sketch.* Consider the energy functions

$$\begin{aligned}
\mathcal{E}^F(t) &= b(t)(f(X(t)) - f(x_*)) + c(t)\left\|X(t) - x_*\right\|^2 + \dot{b}(t)\left\langle X(t) - x_*, \dot{X}(t)\right\rangle + \frac{b(t)}{2}\left\|\dot{X}(t)\right\|^2, \\
\mathcal{E}^G(t) &= \frac{f(Y(t)) - f(Y(T))}{b(T-t)} - c(T-t)\left\|\frac{Y(t) - Y(T)}{b(T-t)}\right\|^2 + \frac{b(T-t)}{2}\left\|\frac{d}{dt}\left\{\frac{Y(t) - Y(T)}{b(T-t)}\right\}\right\|^2.
\end{aligned}$$

By applying ($\mu$-strong) convexity inequalities and performing the calculation, one will obtain the following inequalities:

$$\dot{\mathcal{E}}^F(t) \leq (\ldots), \qquad \dot{\mathcal{E}}^G(T-t) \leq (\ldots).$$

The right-hand sides of these two inequalities are the same under the identification $X(t) - x_* = (Y(T-t) - Y(T))/b(t)$. □

Recall the observation in [5] (available on my homepage) that the PEP kernel can be obtained by integrating the kernel expression for the right-hand side of the inequality $\dot{\mathcal{E}}(t) \leq (...)$ and then

adding some kernels for correction. In the previously known proof of the equivalence between the two PEP analyses (see [4, Thm. 2] or [7, Thm. 3]), the PEP kernels are the anti-transposes of each other, but there is no such relationship between the integrands $\dot{\mathcal{E}}^F(t)$ and $\dot{\mathcal{E}}^G(t)$. If the known proof is not the "right" one, then in the "right" proof, one might hope that the right-hand sides of $\dot{\mathcal{E}}^F(t) \le (\dots)$ and $\dot{\mathcal{E}}^G(T-t) \le (\dots)$ are the anti-transposes of each other. This leads to the identification

$$X(t) - x_* = \frac{Y(T-t) - Y(T)}{\lambda^F(t)},$$

suggesting that the two PEPs should be formulated in the $\{X(t)\}$ and $\{Y(t)\}$ bases (or their variants).

**Remark 1.** *In [1], Theorem 1 was stated in a more general form by allowing the coefficient of $\nabla f(X)$ to be an arbitrary function. I think this generalization can also be handled through the time-dilation argument in [12], but I have not checked this in detail.*

*The essence of Theorem 1 is not the specific energy functions $\mathcal{E}^F(t)$ and $\mathcal{E}^G(t)$. Instead, it is about making the right-hand sides of the two inequalities $\dot{\mathcal{E}}^F(t) \le (\dots)$ and $\dot{\mathcal{E}}^G(T-t) \le (\dots)$ equal under the identification $X(t) - x_* = (Y(T-t) - Y(T))/\lambda^F(t)$.*

## 2.3 Evidence 3: Computation of PEP kernels

First, look at the expressions for the two PEP kernels in [7]:

$$S^F(t,\tau) = \nu\left(\lambda^F(t)H^F(t,\tau) + \dot{\lambda}^F(t)\int_\tau^t H^F(s,\tau)\,ds\right) - 2\alpha^F(t)\alpha^F(\tau), \quad t \ge \tau,$$

and

$$S^G(t,\tau) = \nu\bar{H}^G(t,\tau) - 2\alpha^G(t)\alpha^G(\tau), \quad t \ge \tau.$$

Note that $S^G(t,\tau)$ is concisely written using the kernel $\bar{H}^G(t,\tau)$, which represents the operator mapping $\nabla\hat{f}_t(Y(t))$ to $-\dot{Y}(t)$. There was an impression that $S^F(t,\tau)$ could be written equally concisely using the "right" H-kernel $\bar{H}^F(t,\tau)$, which should be given by

$$\bar{H}^F(t,\tau) = \lambda^F(t)H^F(t,\tau) + \dot{\lambda}^F(t)\int_\tau^t H^F(s,\tau)\,ds.$$

It follows that

$$-\int_0^t \bar{H}^F(t,\tau)\nabla f(X(\tau))\,d\tau = \lambda^F(t)\dot{X}(t) + \dot{\lambda}^F(t)\,(X(t)-x_0) = \frac{d}{dt}\left\{\lambda^F(t)\,(X(t)-x_0)\right\}.$$

Thus, $\bar{H}^F(t,\tau)$ represents the operator mapping $\nabla f(X(t))$ to $-\dot{W}(t)$, where $W(t) = \lambda^F(t)(X(t) - x_0)$.

Recall that $\bar{H}^F$ and $\bar{H}^G$ are the anti-transposes of each other (which, together with $\alpha^F(t) = \alpha^G(T-t)$, implies that the PEP kernels are the anti-transposes of each other, in the previously known proof). If the terms $-2\alpha^F(t)\alpha^F(\tau)$ and $-2\alpha^G(t)\alpha^G(\tau)$ are ignored, the PEP kernels are simply the H-kernels (with modified outputs $\dot{W}(t)$ or modified inputs $\nabla\hat{f}_t(Y(t))$). By using the bases $\{\dot{W}(t)\}$ and $\{\dot{Y}(t)\}$, the PEP kernels[1] become $(\bar{H}^F)^{-1}$ and $(\bar{H}^G)^{-1}$. Once one has checked $(\mathrm{H}^{-1})^{\mathrm{A}} = (\mathrm{H}^{\mathrm{A}})^{-1}$, it is immediate to see that these are the anti-transposes of each other.

[1] It may be an operator which cannot be represented by an integral kernel. Thus, "PEP operator" and "H-operator" would be more appropriate.

# 3 Can one go further?

Inspired by the findings for gradient descent [11, Table 2] and optimal methods [9, Appendix E.2], [3, Table 3], one might conjecture the following six dualities:

| Performance criterion | | Performance criterion |
|---|---|---|
| $(f(x_N) - f_*)/\|x_0 - x_*\|^2$ | $\longleftrightarrow$ | $\|\nabla f(x_N)\|^2/(f(x_0) - f_*)$ |
| $\|x_N - x_*\|^2/\|x_0 - x_*\|^2$ | $\longleftrightarrow$ | $\|\nabla f(x_N)\|^2/\|\nabla f(x_0)\|^2$ |
| $\|x_N - x_*\|^2/(f(x_0) - f_*)$ | $\longleftrightarrow$ | $(f(x_N) - f_*)/\|\nabla f(x_0)\|^2$ |
| $\|\nabla f(x_N)\|^2/\|x_0 - x_*\|^2$ | $\longleftrightarrow$ | $\|\nabla f(x_N)\|^2/\|x_0 - x_*\|^2$ |
| $(f(x_N) - f_*)/(f(x_0) - f_*)$ | $\longleftrightarrow$ | $(f(x_N) - f_*)/(f(x_0) - f_*)$ |
| $\|x_N - x_*\|^2/\|\nabla f(x_0)\|^2$ | $\longleftrightarrow$ | $\|x_N - x_*\|^2/\|\nabla f(x_0)\|^2$ |

In particular, the correspondence between the performance criterion $\|\nabla f(x_N)\|^2/\|x_0 - x_*\|^2$ and itself was studied in the context of monotone inclusion (or equivalently, fixed-point iteration) in [13, Thm. 5.2]. One might try to find a viewpoint that allows us to see the six dualities in a unified vision instead of treating them separately.

The theory in [6, §3] (or its variant) cannot be used to prove that the exact convergence rate of $\mathbf{H}^F$ always correspond to the exact convergence rate of $\mathbf{H}^G$ (and my numerical experiments suggest this is indeed not the case). But the correspondence seems to hold for gradient descent and optimal methods (I have not checked all cases). One might try to provide a theory to explain this phenomenon.

After writing [6], I have not worked on this topic. I think other researchers will go (or already have gone) much further than me.

# 4 Errata

- Equation 3.3: Add $\frac{\tau}{2}\|x_0 - x_*^+\|^2$ to the RHS.

# References


[1] Jaeyeon Kim (assisted by others). An unpublished result, 2022.

[2] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1-2):451–482, 2014.

[3] Shuvomoy Das Gupta, Bart PG Van Parys, and Ernest K Ryu. Branch-and-bound performance estimation programming: a unified methodology for constructing optimal optimization methods. *Mathematical Programming*, 204:567–639, 2024.

[4] Jaeyeon Kim, Asuman Ozdaglar, Chanwoo Park, and Ernest Ryu. Time-reversed dissipation induces duality between minimizing gradient norm and function value. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Jungbin Kim. Notes to "Convergence analysis of ODE models for accelerated first-order methods via positive semidefinite kernels" by Jungbin Kim and Insoon Yang.

[6] Jungbin Kim. A proof of the exact convergence rate of gradient descent. *arXiv preprint arXiv:2412.04427*, 2024.

[7] Jungbin Kim and Insoon Yang. Convergence analysis of ODE models for accelerated first-order methods via positive semidefinite kernels. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Jaewook J Suh, Gyumin Roh, and Ernest K Ryu. Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems. In *International Conference on Machine Learning*, pages 20640–20667. PMLR, 2022.

[9] Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, pages 1–38, 2022.

[10] Adrien B. Taylor, Julien M. Hendrickx, and Fran¸cois Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Math. Program.*, 161(1-2):307–345, 2017.

[11] Adrien B. Taylor, Julien M. Hendrickx, and Fran¸cois Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *J. Optim. Theory Appl.*, 178(2):455–476, 2018.

[12] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[13] Taeho Yoon, Jaeyeon Kim, Jaewook J Suh, and Ernest K Ryu. Optimal acceleration for minimax and fixed-point problems is not unique. In *International Conference on Machine Learning*. PMLR, 2024.