10 Acceleration

Goal

Heavy-ball, momentum, accelerated proximal gradient, FISTA, optimal rate of convergence

Alert 10.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 10.2: Problem

We revisit the following problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad \ell(\mathbf{w}) + r(\mathbf{w}), \tag{10.1}$$

where ℓ is an $\mathsf{L} = \mathsf{L}^{[1]}$-smooth function (w.r.t. $\|\cdot\|_2$) and r is any function whose proximal map is well-defined and easily computable. We solved (10.1) in Lecture 4 using the proximal gradient Algorithm 4.17, and obtained the $O(1/t)$ rate of convergence (in terms of function value) when both ℓ and r are convex. In this lecture we improve the convergence rate to $O(1/t^2)$, which is optimal.

Algorithm 10.3: Heavy ball (Polyak 1964)

To motivate our development, let us first consider the simpler unconstrained minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad f(\mathbf{w}),$$

where f is $\mathsf{L}^{[1]}$-smooth. We have seen the gradient Algorithm 2.4 may zigzag when f is ill-conditioned (i.e. the ratio between the largest and smallest eigenvalues of $\nabla^2 f$ is large). To address this issue, Polyak (1964) proposed the following heavy-ball method:

$$\mathbf{w}_{t+1} = \underbrace{\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)}_{\text{gradient step}} + \underbrace{\beta_t(\mathbf{w}_t - \mathbf{w}_{t-1})}_{\text{momentum}} = \underbrace{(1+\beta_t)\mathbf{w}_t - \beta_t \mathbf{w}_{t-1}}_{\text{extrapolation}} - \eta_t \nabla f(\mathbf{w}_t), \tag{10.2}$$

where typically $\mathbf{w}_1 = \mathbf{w}_0$ (so that at $t = 1$ we start with the usual gradient step).

To see the physical meaning of (10.2), let us derive the underlying continuous analogue:

$$\begin{aligned}\mathbf{0} &= [(\mathbf{w}_{t+1} - \mathbf{w}_t) - (\mathbf{w}_t - \mathbf{w}_{t-1})] + (1-\beta_t)(\mathbf{w}_t - \mathbf{w}_{t-1}) + \eta_t \nabla f(\mathbf{w}_t) \\ &\approx \ddot{\mathbf{w}}(t) + (1-\beta_t)\dot{\mathbf{w}}(t) + \eta_t \nabla f(\mathbf{w}(t)),\end{aligned}$$

which follows from the usual finite-difference approximation of the time derivative $\dot{\mathbf{w}}$. We may now interpret $\mathbf{w}(t)$ as the position of a heavy ball, whose velocity is $\dot{\mathbf{w}}(t)$ and momentum is $\ddot{\mathbf{w}}(t)$ whereas the function f acts as its potential energy. From the last equation in (10.2) we also see the extrapolation effect if $\beta_t > 0$ (as opposed to interpolation when $\beta_t < 0$, which amounts to a convex combination of the previous two positions $\mathbf{w}_{t-1}$ and $\mathbf{w}_t$).

With suitable choices of η_t and β_t, heavy-ball was shown to converge optimally on strongly convex quadratic functions. However, proving its convergence rate for even $\mathsf{L}^{[1]}$-smooth functions has remained challenging.

Polyak, B. T. (1964). "Some methods of speeding up the convergence of iteration methods". *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 791–803.

Remark 10.4: Convergence of heavy ball

Danilova et al. (2020) proved the convergence of the heavy ball algorithm when ℓ is $\mathsf{L}^{[1]}$-smooth, by constructing an interesting Lyapunov function. See also Ghadimi et al. (2015).

Danilova, M., A. Kulakova, and B. Polyak (2020). "Non-monotone Behavior of the Heavy Ball Method". In: *Difference Equations and Discrete Dynamical Systems with Applications*. Ed. by M. Bohner, S. Siegmund, R. Šimon Hilscher, and P. Stehlík, pp. 213–230.

Ghadimi, E., H. R. Feyzmahdavian, and M. Johansson (2015). "Global convergence of the Heavy-ball method for convex optimization". In: *European Control Conference (ECC)*, pp. 310–315.

Algorithm 10.5: Nesterov's momentum

In heavy ball, we perform an extrapolation step and then a gradient step. However, the gradient is not the computed at the extrapolated position but the current position $\mathbf{w}_t$. This motivates us to consider the following ingenious variation due to Nesterov (1983):

$$\begin{aligned}\mathbf{z}_{t+1} &= \mathbf{w}_t + \beta_t(\mathbf{w}_t - \mathbf{w}_{t-1})\\ \mathbf{w}_{t+1} &= \mathbf{z}_{t+1} - \eta_t \nabla f(\mathbf{z}_{t+1}),\end{aligned}$$

where the only difference from heavy ball (10.2) is that the gradient is now evaluated at the extrapolated position $\mathbf{z}_{t+1}$. This modification may seem minor *in retrospective*, but quite remarkably it leads to an optimal convergence rate $O(1/t^2)$.

A similar continuous analogue was derived by Su et al. (2016). Indeed, with a constant step size $\eta_t \equiv \eta$ and set $\mathbf{w}(t) = \mathbf{w}_{t/\sqrt{\eta}}$ we obtain as before:

$$\begin{aligned}\mathbf{0} &= \frac{(\mathbf{w}_{t/\sqrt{\eta}+1} - \mathbf{w}_{t/\sqrt{\eta}}) - (\mathbf{w}_{t/\sqrt{\eta}} - \mathbf{w}_{t/\sqrt{\eta}-1})}{\sqrt{\eta}} + (1-\beta_{t/\sqrt{\eta}})\frac{\mathbf{w}_{t/\sqrt{\eta}} - \mathbf{w}_{t/\sqrt{\eta}-1}}{\sqrt{\eta}} + \sqrt{\eta}\nabla f(\mathbf{z}_{t/\sqrt{\eta}+1})\\ &= \frac{(\mathbf{w}(t+\sqrt{\eta}) - \mathbf{w}(t)) - (\mathbf{w}(t) - \mathbf{w}(t-\sqrt{\eta}))}{\sqrt{\eta}} + (1-\beta(t))\frac{\mathbf{w}(t) - \mathbf{w}(t-\sqrt{\eta})}{\sqrt{\eta}} + \sqrt{\eta}\nabla f(\mathbf{z}(t+\sqrt{\eta}))\\ &= \ddot{\mathbf{w}}(t)\sqrt{\eta} + (1-\beta(t))[\dot{\mathbf{w}}(t) - \tfrac{1}{2}\ddot{\mathbf{w}}(t)\sqrt{\eta}] + \sqrt{\eta}\nabla f\big(\mathbf{w}(t) + \beta(t)(\mathbf{w}(t) - \mathbf{w}(t-\sqrt{\eta}))\big) + o(\sqrt{\eta})\\ &= \sqrt{\eta}\Big[\frac{\beta(t)+1}{2}\ddot{\mathbf{w}}(t) + \frac{1-\beta(t)}{\sqrt{\eta}}\dot{\mathbf{w}}(t) + \nabla f\big(\mathbf{w}(t)\big) + o(1) + O(\sqrt{\eta}\beta(t))\Big].\end{aligned}$$

Letting

$$\beta(t) = \beta_{t/\sqrt{\eta}} = \frac{t - \sqrt{\eta}\frac{a-1}{2}}{t + \sqrt{\eta}\frac{a+1}{2}}, \quad a \geq 3$$

and letting $\eta \to 0$ we have

$$\frac{\beta(t)+1}{2} = 1 + o(1), \quad \frac{1-\beta(t)}{\sqrt{\eta}} = \frac{a}{t} + o(1), \quad \sqrt{\eta}\beta(t) = o(1).$$

Dropping the lower order term $o(1)$ we finally arrive at:

$$\ddot{\mathbf{w}}(t) + \frac{a}{t}\dot{\mathbf{w}}(t) + \nabla f\big(\mathbf{w}(t)\big) = \mathbf{0},$$

which has been heavily studied since.

Nesterov, Y. E. (1983). "A Method for Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$". *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376.

Su, W., S. Boyd, and E. J. Candès (2016). "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43.

Algorithm 10.6: FISTA (Beck and Teboulle 2009; Nesterov 2013)

We now extend Nesterov's momentum to the composite problem (10.1). We simply keep the extrapolation step but augment the gradient step, which amounts to minimizing the familiar quadratic upper bound:

$$\begin{aligned}\mathbf{w}_t &= \operatorname*{argmin}_{\mathbf{w}} \ \ell(\mathbf{z}_t) + \langle \mathbf{w} - \mathbf{z}_t, \nabla \ell(\mathbf{z}_t)\rangle + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{z}_t\|_2^2 + r(\mathbf{w}) \\ &= \mathrm{P}_r^{\eta_t}(\mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)).\end{aligned}$$

Algorithm: Accelerated Proximal Gradient, a.k.a. FISTA

Input: $\mathbf{w}_0 = \mathbf{z}_1, \gamma_1 = 1, \eta_0$

1 **for** $t = 1, 2, \ldots$ **do**

2 choose step size $\eta_t \le \eta_{t-1}$ `// step size can only decrease`

3 $\mathbf{u}_t = \mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)$ `// gradient step w.r.t.` ℓ

4 $\mathbf{w}_t = \mathrm{P}_r^{\eta_t}(\mathbf{u}_t) = \operatorname{argmin}_{\mathbf{u}} \frac{1}{2\eta_t}\|\mathbf{u}_t - \mathbf{u}\|_2^2 + r(\mathbf{u})$ `// proximal step w.r.t.` r

5 $\gamma_{t+1} = \frac{1+\sqrt{1+4\gamma_t^2}}{2}$

6 $\beta_t = \frac{\gamma_t - 1}{\gamma_{t+1}}$ `// momentum size`

7 $\mathbf{z}_{t+1} = \mathbf{w}_t + \beta_t(\mathbf{w}_t - \mathbf{w}_{t-1})$ `// extrapolation`

Beck, A. and M. Teboulle (2009). "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202.

Nesterov, Y. E. (2013). "Gradient Methods for Minimizing Composite Functions". *Mathematical Programming, Series B*, vol. 140, pp. 125–161.

Remark 10.7: Observations on FISTA

A few comments on FISTA are in order:

- Needless to say, when $r \equiv 0$ in Line 7, FISTA reduces to the original algorithm of Nesterov.
- With the choice $\gamma_1 = 1$, $\mathbf{w}_0$ does not really play any role: the first step of the algorithm (i.e. $t = 1$) is simply a proximal gradient step.
- FISTA in Line 7 requires the smooth function ℓ to be defined over the entire space $\mathbb{R}^d$, since the extrapolated sequence $\mathbf{z}_t$ may jump outside of $\operatorname{dom} r$. On the other hand, the proximal sequence $\mathbf{w}_t$ remains in $\operatorname{dom} r$ by construction. Variants that make sure $\mathbf{z}_t$ remains in $\operatorname{dom} r$ include (Nesterov 2005; Auslender and Teboulle 2006).
- We note that the momentum choice $\beta_t = \frac{\gamma_t - 1}{\gamma_{t+1}}$ is w.l.o.g. Indeed, given any sequence β_t, we may recover $\gamma_t = 1 \iff \beta_t = 0$ (which corresponds to a standard gradient step), and given γ_j such that $\beta_{j-1} = 0$ and for all $t \in [j, i], \beta_t \neq 0$, we have

$$\forall t \in [j, i], \quad \gamma_{t+1} = \frac{\gamma_t - 1}{\beta_t} = \frac{\gamma_j - 1 - \sum_{m=j}^{t-1} \prod_{k=j}^{m} \beta_k}{\prod_{k=j}^{t} \beta_k}.$$

  In particular, the choice

$$\gamma_t = \frac{t + a - 2}{a - 1}, \quad \text{or equivalently} \quad \beta_t = \frac{t - 1}{t + a - 1}, \quad a \ge 3, \tag{10.3}$$

  works equally well. In fact, for $a > 3$, Chambolle and Dossal (2015) and Attouch and Peypouquet (2016) proved that both the function value and the proximal iterate $\mathbf{w}_t$ converges at $o(1/t^2)$.

Nesterov, Y. E. (2005). "Smooth Minimization of Non-Smooth Functions". *Mathematical Programming, Series A*, vol. 103, pp. 127–152.

Auslender, A. and M. Teboulle (2006). "Interior Gradient and Proximal Methods for Convex and Conic Optimization". *SIAM Journal on Optimization*, vol. 16, no. 3, pp. 697–725.
Chambolle, A. and C. Dossal (2015). "On the Convergence of the Iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm"". *Journal of Optimization Theory and Application*, vol. 166, pp. 968–982.
Attouch, H. and J. Peypouquet (2016). "The Rate of Convergence of Nesterov's Accelerated Forward-Backward Method is Actually Faster Than $1/k^2$". *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1824–1834.

Theorem 10.8: Accelerated Proximal Gradient (Beck and Teboulle 2009; Nesterov 2013)

Suppose $\ell : \mathbb{R}^d \to \mathbb{R}$ is $\mathsf{L}^{[1]}$-smooth and convex, $r : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and convex, and $\eta_t \equiv \eta \le 1/\mathsf{L}^{[1]}$. Then, the proximal sequence $\{\mathbf{w}_t\}$ generated by Line 7 satisfies: for all $\mathbf{w}$ and $t \ge 1$,

$$f(\mathbf{w}_t) \le f(\mathbf{w}) + \frac{\|\mathbf{w} - \mathbf{z}_1\|_2^2}{2\eta_t \gamma_t^2} \le f(\mathbf{w}) + \frac{2\|\mathbf{w} - \mathbf{z}_1\|_2^2}{\eta_t (t+1)^2}. \tag{10.4}$$

Proof: We learned the following inspiring proof from Tseng (2010). It follows from the step size choice, $\mathsf{L}^{[1]}$-smoothness, and composite optimality Proposition 4.20 that for any $\mathbf{v}$,

$$\begin{aligned} f(\mathbf{w}_t) &\le r(\mathbf{w}_t) + \ell(\mathbf{z}_t) + \langle \mathbf{w}_t - \mathbf{z}_t, \nabla \ell(\mathbf{z}_t) \rangle + \tfrac{1}{2\eta_t} \|\mathbf{w}_t - \mathbf{z}_t\|_2^2 \\ &\le r(\mathbf{v}) + \ell(\mathbf{z}_t) + \langle \mathbf{v} - \mathbf{z}_t, \nabla \ell(\mathbf{z}_t) \rangle + \tfrac{1}{2\eta_t} \|\mathbf{v} - \mathbf{z}_t\|_2^2 - \tfrac{1}{2\eta_t} \|\mathbf{v} - \mathbf{w}_t\|_2^2 \\ &\le r(\mathbf{v}) + \ell(\mathbf{v}) + \tfrac{1}{2\eta_t} \|\mathbf{v} - \mathbf{z}_t\|_2^2 - \tfrac{1}{2\eta_t} \|\mathbf{v} - \mathbf{w}_t\|_2^2, \end{aligned} \tag{10.5}$$

where the last inequality is due to the convexity of ℓ. Thus far, everything is the same as in the proof of Theorem 4.21.

Now to get a faster convergence, we need to somehow make the quadratic terms above diminish faster, which is achieved by choosing the convex combination $\mathbf{v}_t := (1 - \frac{1}{\gamma_t})\mathbf{w}_{t-1} + \frac{1}{\gamma_t}\mathbf{w}$ for some arbitrary $\mathbf{w} \in \operatorname{dom} f$. Plugging into (10.5) we obtain

$$\begin{aligned} f(\mathbf{w}_t) &\le f(\mathbf{v}_t) + \tfrac{1}{2\eta_t} \|\mathbf{v}_t - \mathbf{z}_t\|_2^2 - \tfrac{1}{2\eta_t} \|\mathbf{v}_t - \mathbf{w}_t\|_2^2 \\ &\le (1 - \tfrac{1}{\gamma_t}) f(\mathbf{w}_{t-1}) + \tfrac{1}{\gamma_t} f(\mathbf{w}) + \tfrac{1}{2\eta_t \gamma_t^2} \left[\|\mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \mathbf{z}_t)\|_2^2 - \|\mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \mathbf{w}_t)\|_2^2 \right]. \end{aligned}$$

Define $\mathbf{q}_t = \gamma_{t+1}(\mathbf{z}_{t+1} - \mathbf{w}_t) = (\gamma_t - 1)(\mathbf{w}_t - \mathbf{w}_{t-1})$, since $\beta_t = \frac{\gamma_t - 1}{\gamma_{t+1}}$. We verify

$$\mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \mathbf{z}_t) = \mathbf{w} - \mathbf{w}_{t-1} - \mathbf{q}_{t-1}, \quad \mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \mathbf{w}_t) = \mathbf{w} - \mathbf{w}_t - \mathbf{q}_t,$$

and thus

$$f(\mathbf{w}_t) - f(\mathbf{w}) \le (1 - \tfrac{1}{\gamma_t})[f(\mathbf{w}_{t-1}) - f(\mathbf{w})] + \tfrac{1}{2\eta_t \gamma_t^2} \left[\|\mathbf{w} - \mathbf{w}_{t-1} - \mathbf{q}_{t-1}\|_2^2 - \|\mathbf{w} - \mathbf{q}_t - \mathbf{w}_t\|_2^2 \right].$$

Using the relation $\gamma_{t-1}^2 = \gamma_t^2 - \gamma_t$, we obtain the recursion

$$\begin{aligned} \eta_t \gamma_t^2 [f(\mathbf{w}_t) - f(\mathbf{w})] + \tfrac{1}{2} \|\mathbf{w} - \mathbf{w}_t - \mathbf{q}_t\|_2^2 &\le \eta_t \gamma_{t-1}^2 [f(\mathbf{w}_{t-1}) - f(\mathbf{w})] + \tfrac{1}{2} \|\mathbf{w} - \mathbf{w}_{t-1} - \mathbf{q}_{t-1}\|_2^2, \\ &\le \eta_{t-1} \gamma_{t-1}^2 [f(\mathbf{w}_{t-1}) - f(\mathbf{w})] + \tfrac{1}{2} \|\mathbf{w} - \mathbf{w}_{t-1} - \mathbf{q}_{t-1}\|_2^2, \end{aligned}$$

using the assumption $\eta_t = \eta_{t-1}$. Telescoping yields

$$\tfrac{1}{2} \|\mathbf{w} - \mathbf{w}_t - \mathbf{q}_t\|_2^2 + \eta_t \gamma_t^2 [f(\mathbf{w}_t) - f(\mathbf{w})] \le \eta_1 \gamma_1^2 [f(\mathbf{w}_1) - f(\mathbf{w})] + \tfrac{1}{2} \|\mathbf{w} - \mathbf{w}_1 - \mathbf{q}_1\|_2^2$$

Since $\gamma_1 = 1$, $\mathbf{q}_1 = \mathbf{0}$ and $\mathbf{w}_1$ is simply a proximal gradient step from $\mathbf{z}_1$. Setting $\mathbf{v} = \mathbf{w}$ and $t = 1$ in (10.5) we further have

$$\eta_t \gamma_t^2 [f(\mathbf{w}_t) - f(\mathbf{w})] \le \eta_1 [f(\mathbf{w}_1) - f(\mathbf{w}) + \tfrac{1}{2\eta_1} \|\mathbf{w} - \mathbf{w}_1\|_2^2] \le \tfrac{1}{2} \|\mathbf{w} - \mathbf{z}_1\|_2^2.$$

Finally, we examine the sequence γ_t. By definition:

$$\frac{1}{2}+\gamma_t \le \frac{1+2\gamma_t}{2} \le \gamma_{t+1} \le \frac{1+\sqrt{1+4\gamma_t^2}}{2} \le \frac{1+1+2\gamma_t}{2} = 1+\gamma_t,$$

implying $t-1+\gamma_1 \ge \gamma_t \ge \frac{t-1}{2}+\gamma_1$. Applying the lower bound on γ_t with $\gamma_1 = 1$ completes the proof. ■

Beck, A. and M. Teboulle (2009). "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202.
Nesterov, Y. E. (2013). "Gradient Methods for Minimizing Composite Functions". *Mathematical Programming, Series B*, vol. 140, pp. 125–161.
Tseng, P. (2010). "Approximation Accuracy, Gradient Methods, and Error Bound for Structured Convex Optimization". *Mathematical Programming, Series B*, vol. 125, pp. 263–295.

Exercise 10.9: Some refinements

If we choose $\mathbf{w} \in \operatorname{argmin} f$, then we can make the following refinements:

- The extrapolation constants need only satisfy

$$\gamma_{t-1}^2 \ge \gamma_t^2 - \gamma_t.$$

  In particular, the choice for γ_t in (10.3) works and enjoys the same bound in Theorem 10.8 (with slightly worse constants).

- We can use Amijo's rule to adaptively choose η_t. However, the condition $\eta_t \le \eta_{t-1}$ needs to be respected, meaning that each Amijo step should start with the step size from the previous iteration.

Algorithm 10.10: Enforcing monotonicity (Beck and Teboulle 2009)

Algorithm: Monotonic FISTA

Input: $\mathbf{w}_0 = \mathbf{z}_1, \gamma_1 = 1, \eta_0$

1 **for** $t = 1, 2, \dots$ **do**
2 choose step size $\eta_t \le \eta_{t-1}$ `// step size can only decrease`
3 $\mathbf{u}_t = \mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)$ `// gradient step w.r.t.` ℓ
4 $\tilde{\mathbf{w}}_t = \mathrm{P}_r^{\eta_t}(\mathbf{u}_t) = \operatorname{argmin}_{\mathbf{u}} \frac{1}{2\eta_t}\|\mathbf{u}_t - \mathbf{u}\|_2^2 + r(\mathbf{u})$ `// proximal step w.r.t.` r
5 choose $\mathbf{w}_t$ such that $f(\mathbf{w}_t) \le f(\tilde{\mathbf{w}}_t)$ `// local improvment`
6 $\gamma_{t+1} = \frac{1+\sqrt{1+4\gamma_t^2}}{2}$
7 $\mathbf{z}_{t+1} = \mathbf{w}_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{\gamma_t}{\gamma_{t+1}}(\tilde{\mathbf{w}}_t - \mathbf{w}_t)$ `// extrapolation`

Unlike the proximal gradient Algorithm 4.17, the objective value $f(\mathbf{w}_t)$ of the accelerated Line 7 may not be monotonically decreasing. However, as noted by Beck and Teboulle (2009), this can be fixed through roll back, i.e., setting $\mathbf{w}_t = \mathbf{w}_{t-1}$ in Algorithm 10.10 whenever jumps happen, i.e. $f(\tilde{\mathbf{w}}_t) > f(\mathbf{w}_{t-1})$. However, we also need to make some adjustment to the extrapolation step, as indicated in the last term on line 7. Indeed, this slight modification allows us to prove the same complexity bound as in Theorem 10.8: Recall,

$$f(\tilde{\mathbf{w}}_t) \le (1-\tfrac{1}{\gamma_t})f(\mathbf{w}_{t-1}) + \tfrac{1}{\gamma_t}f(\mathbf{w}) + \tfrac{1}{2\eta_t\gamma_t^2}\left[\|\mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \mathbf{z}_t)\|_2^2 - \|\mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \tilde{\mathbf{w}}_t)\|_2^2\right],$$

Define $\mathbf{q}_t = \gamma_{t+1}(\mathbf{z}_{t+1} - \mathbf{w}_t) = (\gamma_t - 1)(\mathbf{w}_t - \mathbf{w}_{t-1}) + \gamma_t(\tilde{\mathbf{w}}_t - \mathbf{w}_t) = \mathbf{w}_{t-1} - \gamma_t(\mathbf{w}_{t-1} - \tilde{\mathbf{w}}_t) - \mathbf{w}_t$ and we verify:

$$\mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \mathbf{z}_t) = \mathbf{w} - \mathbf{w}_{t-1} - \mathbf{q}_{t-1}, \quad \mathbf{w} - \mathbf{w}_{t-1} + \gamma_t(\mathbf{w}_{t-1} - \tilde{\mathbf{w}}_t) = \mathbf{w} - \mathbf{w}_t - \mathbf{q}_t.$$

Since $f(\mathbf{w}_t) \le f(\tilde{\mathbf{w}}_t)$ we may continue the rest of the proof as in Theorem 10.8.

Beck, A. and M. Teboulle (2009). "Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems". *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2419–2434.

Remark 10.11: Adaptive restarting

When jumps happen, i.e. $f(\mathbf{w}_t) \geq f(\mathbf{w}_{t-1})$ in Line 7, one may simply reset $\mathbf{w}_t$ to $\mathbf{w}_{t-1}$. Without any amendment in the extrapolation step (such as the one in Algorithm 10.10), the next iteration amounts to a standard proximal gradient step which we know from Theorem 4.21 will decrease the objective value. This heuristic has two issues:

- It is not clear how we can recover the same convergence rate as in Theorem 10.8 (cf. Algorithm 10.10, where we amend the extrapolation step);
- With $\gamma_1 = 1$ the first iteration of Line 7 amounts to a standard proximal gradient step. When jumps happen, the reset also performs a standard proximal gradient step. However, the momentum parameter γ_t continues to grow.

Thus, a natural alternative is to *completely restart* the algorithm by setting $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1}$ and also $\gamma_{t+1} \leftarrow 1$, which appears to accelerate convergence (O'Donoghue and Candès 2015). Kim and Fessler (2018) also studied the restarting trick for the optimized variant in Algorithm 10.12.

O'Donoghue, B. and E. Candès (2015). "Adaptive Restart for Accelerated Gradient Schemes". *Foundations of Computational Mathematics*, vol. 15, pp. 715–732.
Kim, D. and J. A. Fessler (2018). "Adaptive Restart of the Optimized Gradient Method for Convex Optimization". *Journal of Optimization Theory and Applications*, vol. 178, pp. 240–263.

Algorithm 10.12: Optimized gradient descent

Algorithm: Optimized gradient descent

Input: $\mathbf{w}_0 = \mathbf{z}_1, \gamma_1 = 1, \eta_0$

1 **for** $t = 1, 2, \ldots, T$ **do**

2 choose step size $\eta_t \leq \eta_{t-1}$ `// step size can only decrease`

3 $\mathbf{w}_t = \mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)$ `// gradient step w.r.t.` ℓ

4 $\gamma_{t+1} = \frac{1+\sqrt{1+4\gamma_t^2}}{2}$

5 **if** $t = T$ **then**

6 $\gamma_{t+1} = \frac{1+\sqrt{1+8\gamma_t^2}}{2}$

7 $\mathbf{z}_{t+1} = \mathbf{w}_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{\gamma_t}{\gamma_{t+1}}(\mathbf{w}_t - \mathbf{z}_t)$ `// extrapolation`

By refining the performance estimation problem, Kim and Fessler (2016) proposed the optimized variant above (with $r \equiv 0$) and proved the same convergence rate for the extrapolated sequence but with improved constants:

$$f(\mathbf{z}_{T+1}) - f_\star \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{2\eta\gamma_{T+1}^2} \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{\eta(T+1)(T+1+\sqrt{2})}, \quad \eta_t \equiv \eta \leq 1/\mathrm{L}^{[1]},$$

which amounts to a factor of 2 improvement compared to Theorem 10.8 and is known to be tight (Drori 2017). Moreover, Kim and Fessler (2016) also proved the convergence rate (10.4) for the extrapolated sequence $\mathbf{z}_t$ of FISTA in Line 7.

Later, Kim and Fessler (2017) proved the proximal sequence $\mathbf{w}_t$ of the optimized gradient Algorithm 10.12 also converges at a similar rate:

$$f(\mathbf{w}_t) - f_\star \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{4\eta\gamma_t^2} \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{\eta(t+1)^2}.$$

Kim and Fessler (2018a), also Taylor et al. (2017, 2018), extended the optimized variant to the composite setting (i.e. $r \neq 0$), while Kim and Fessler (2018b) optimized the gradient norm instead of the objective values.

Kim, D. and J. A. Fessler (2016). "Optimized first-order methods for smooth convex minimization". *Mathematical Programming*, vol. 159, pp. 81–107.

Drori, Y. (2017). "The exact information-based complexity of smooth convex minimization". *Journal of Complexity*, vol. 39, pp. 1–16.

Kim, D. and J. A. Fessler (2017). "On the Convergence Analysis of the Optimized Gradient Method". *Journal of Optimization Theory and Applications*, vol. 172, pp. 187–205.

— (2018a). "Another Look at the Fast Iterative Shrinkage/Thresholding Algorithm (FISTA)". *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 223–250.

Taylor, A. B., J. M. Hendrickx, and F. Glineur (2017). "Exact Worst-Case Performance of First-Order Methods for Composite Convex Optimization". *SIAM Journal on Optimization*, vol. 27, no. 3, pp. 1283–1313.

— (2018). "Exact Worst-Case Convergence Rates of the Proximal Gradient Method for Composite Convex Minimization". *Journal of Optimization Theory and Applications*, vol. 178, pp. 455–476.

Kim, D. and J. A. Fessler (2018b). "Generalizing the Optimized Gradient Method for Smooth Convex Minimization". *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1920–1950.

Remark 10.13: What if no minimizer?

Bauschke et al. (2019) studied the intriguing setting where FISTA is applied to a minimization problem that does not admit any minimizer.

Bauschke, H. H., M. N. Bui, and X. Wang (2019). "Applying FISTA to optimization problems (with or) without minimizers". *Mathematical Programming*.