Lecture 13: Linear Coupling and Acceleration

Lecture 12 says that the smooth convex first-order benchmark is order $1/T^2$. Lecture 13 proves that this benchmark is achievable. The route is linear coupling [AZO17]: one gradient step gives primal progress, one mirror step gives dual progress, and the coupled query point makes the two forms of progress match.

13.1 Euclidean Preview

Before proving the full coupling inequality, here is the basic Euclidean intuition. At the same query point x, write $g = \nabla f(x)$. A gradient step uses smoothness:

$$y^+ = x - \frac{1}{L} g, \qquad f(y^+) \le f(x) - \frac{1}{2L} \|g\|_2^2 .$$

Thus a large gradient immediately gives objective decrease.

Mirror descent uses the same linear information differently. With the Euclidean mirror map $h(z) = \frac{1}{2}\|z\|_2^2$, a mirror step from a separate point z is

$$z^+ = \arg\min_q \left\{ \alpha \langle g, q - z \rangle + \frac{1}{2} \|q - z\|_2^2 \right\} = z - \alpha g.$$

For any comparator u, this gives the distance-drop inequality

$$\alpha \langle g, z - u \rangle \le \frac{1}{2} \|u - z\|_2^2 - \frac{1}{2} \left\|u - z^+\right\|_2^2 + \frac{\alpha^2}{2} \|g\|_2^2 .$$

So when the gradient is small, the direct descent term is weak, but the linear term $\langle g, z - u \rangle$ can still be converted into telescoping progress in a distance potential.

Acceleration couples these two uses of the same gradient. Gradient descent is good in the large-gradient case; mirror descent is good at turning the remaining linear signal into potential decrease. Linear coupling chooses the query point so the two progress inequalities can be added without an explicit case split.

13.2 Primal and Dual Progress

The proof of acceleration is easier to remember if we separate two one-step progress inequalities. Primal progress comes from the usual smoothness upper model behind a gradient-type step. Dual progress comes from the mirror-step inequality from Lecture 8, written here in a form designed for coupling.

Lemma 13.1 (Primal progress). *Let E be a finite-dimensional normed vector space with norm $\|\cdot\|$, and let $X \subseteq E$ be nonempty. Let f be differentiable on an open convex set containing X,* *and assume that f satisfies the L-smoothness inequality*

$$f(v) \le f(x) + \langle \nabla f(x), v - x\rangle + \frac{L}{2}\|v - x\|^2 \qquad \forall x, v \in X.$$

Fix $x \in X$, set $g := \nabla f(x)$, and choose

$$y^+ \in \operatorname*{arg\,min}_{q \in X} \left\{ \langle g, q - x\rangle + \frac{L}{2}\|q - x\|^2 \right\}.$$

Then

$$f(y^+) \le f(x) + \langle g, v - x\rangle + \frac{L}{2}\|v - x\|^2 \qquad \forall v \in X.$$

Proof of Lemma 13.1. By L-smoothness at x,

$$f(y^+) \le f(x) + \left\langle g, y^+ - x \right\rangle + \frac{L}{2}\left\|y^+ - x\right\|^2.$$

The definition of y^+ makes the last two terms no larger than $\langle g, v - x\rangle + \frac{L}{2}\|v - x\|^2$ for every $v \in X$. □

Lemma 13.2 (Dual progress)**.** *Let E be a finite-dimensional vector space, and let $X \subseteq E$ be nonempty and convex. Let h be convex and differentiable on an open convex set containing X, and write*

$$D_h(a, b) := h(a) - h(b) - \langle \nabla h(b), a - b\rangle \qquad \forall a, b \in X.$$

Fix $z \in X$, a covector $g \in E^*$, and a scalar $\alpha > 0$. Choose

$$z^+ \in \operatorname*{arg\,min}_{q \in X} \left\{ \alpha \langle g, q - z\rangle + D_h(q, z) \right\}.$$

Then, for every $u \in X$,

$$\alpha \left\langle g, z^+ - z \right\rangle + D_h(z^+, z) \le \alpha \langle g, u - z\rangle + D_h(u, z) - D_h(u, z^+).$$

Proof idea of Lemma 13.2. Let

$$M(q) := \alpha \langle g, q - z\rangle + D_h(q, z).$$

The mirror step says that z^+ minimizes M over X. Instead of memorizing the three-point identity, remember the following picture:

$$\text{model value at } z^+ \le \text{linearization of } M \text{ at } z^+ \text{ evaluated at } u.$$

After expanding D_h, that affine linearization is exactly

$$\alpha \langle g, u - z\rangle + D_h(u, z) - D_h(u, z^+).$$

This is exactly the displayed inequality.

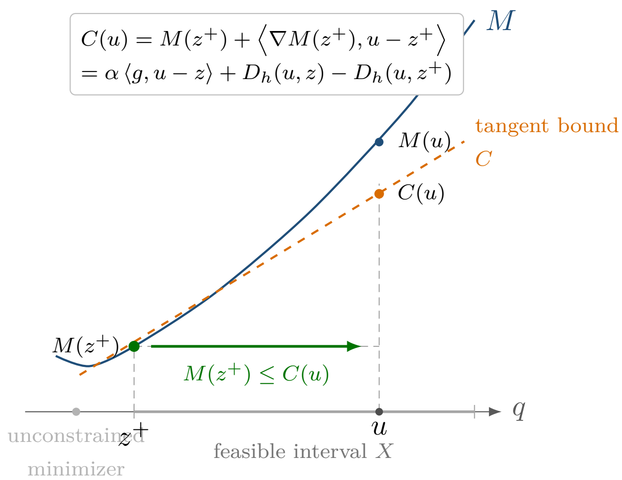


Figure 1: Dual-progress picture. At a constrained minimizer z^+, the tangent need not be flat, but it has nonnegative slope in every feasible direction. Thus the model value at z^+ is at most the tangent bound $C(u)$ at any comparator $u \in X$.

Proof of Lemma 13.2. The first-order condition for minimizing

$$M(q) := \alpha \langle g, q - z\rangle + D_h(q, z)$$

over the convex set X gives

$$\left\langle \alpha g + \nabla h(z^+) - \nabla h(z), u - z^+ \right\rangle \ge 0 \qquad \forall u \in X.$$

Hence

$$\alpha \left\langle g, z^+ - u \right\rangle \le \left\langle \nabla h(z^+) - \nabla h(z), u - z^+ \right\rangle.$$

The Bregman three-point identity gives

$$\left\langle \nabla h(z^+) - \nabla h(z), u - z^+ \right\rangle = D_h(u, z) - D_h(u, z^+) - D_h(z^+, z).$$

Thus

$$\alpha \left\langle g, z^+ - u \right\rangle + D_h(z^+, z) \le D_h(u, z) - D_h(u, z^+).$$

Since $z^+ - u = (z^+ - z) - (u - z)$, this is equivalent to

$$\alpha \left\langle g, z^+ - z \right\rangle + D_h(z^+, z) \le \alpha \langle g, u - z\rangle + D_h(u, z) - D_h(u, z^+).$$

□

13.3 Linear Coupling

Primal progress is naturally centered at the query point x. Dual progress is naturally centered at the mirror point z. Linear coupling chooses x on the segment between the last gradient point y and the mirror point z, so that the mirror displacement at z becomes a scaled gradient-model displacement at x.

Definition 13.1 (Linear coupling)**.** Let E be a finite-dimensional normed vector space with norm $\|\cdot\|$, and let $X \subseteq E$ be nonempty and convex. Let f and h be differentiable on an open convex set containing X, let $L, \rho > 0$, and write

$$D_h(a,b) := h(a) - h(b) - \langle \nabla h(b), a - b \rangle .$$

Given $y_t, z_t \in X$ and $\tau_t \in (0,1]$, define

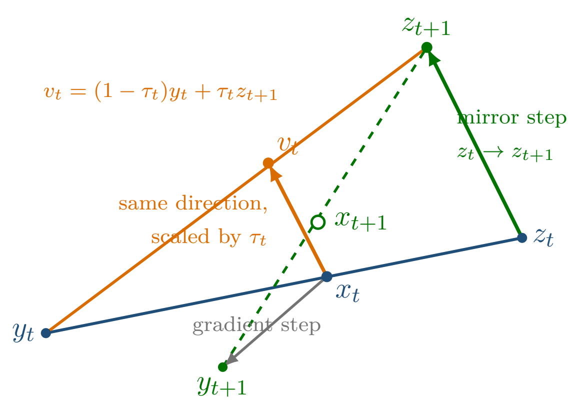
$$x_t := (1-\tau_t)y_t + \tau_t z_t, \qquad g_t := \nabla f(x_t).$$

The linear-coupling step chooses

$$y_{t+1} \in \operatorname*{arg\,min}_{q \in X} \left\{ \langle g_t, q - x_t \rangle + \frac{L}{2} \|q - x_t\|^2 \right\}$$

and

$$z_{t+1} \in \operatorname*{arg\,min}_{q \in X} \left\{ \frac{1}{\tau_t} \langle g_t, q - z_t \rangle + \frac{L}{\rho} D_h(q, z_t) \right\}.$$



$$v_t - x_t = \tau_t(z_{t+1} - z_t), \qquad z_t - z_{t+1} = \frac{x_t - v_t}{\tau_t}.$$

Figure 2: The two-dimensional geometry of the coupling step. The proof uses the auxiliary point v_t, obtained by replacing z_t with z_{t+1} while keeping the same coefficient τ_t; the next actual query x_{t+1} uses the new pair y_{t+1}, z_{t+1} and the next coefficient τ_{t+1}.

Theorem 13.3 (One-step linear-coupling inequality)**.** *Let E be a finite-dimensional normed vector space with norm $\|\cdot\|$, and let $X \subseteq E$ be nonempty and convex. Let f be convex and differentiable on an open convex set containing X, and assume that f satisfies the L-smoothness inequality on X:*

$$f(v) \le f(x) + \langle \nabla f(x), v - x \rangle + \frac{L}{2} \|v - x\|^2 \qquad \forall x, v \in X.$$

Let h be convex and differentiable on an open convex set containing X, write

$$D_h(a,b) := h(a) - h(b) - \langle \nabla h(b), a - b \rangle ,$$

and assume that h satisfies the ρ-strong-convexity inequality

$$D_h(a,b) \geq \frac{\rho}{2} \|a-b\|^2 \qquad \forall a,b \in X.$$

Let $y, z \in X$, $\tau \in (0,1]$, and

$$x := (1-\tau)y + \tau z, \qquad g := \nabla f(x).$$

Choose

$$y^+ \in \arg\min_{q \in X} \left\{ \langle g, q-x \rangle + \frac{L}{2} \|q-x\|^2 \right\},$$

and

$$z^+ \in \arg\min_{q \in X} \left\{ \frac{1}{\tau} \langle g, q-z \rangle + \frac{L}{\rho} D_h(q,z) \right\}.$$

Then, for every $u \in X$,

$$\frac{\rho}{L\tau^2}(f(y^+) - f(u)) + D_h(u,z^+) \leq \frac{\rho(1-\tau)}{L\tau^2}(f(y) - f(u)) + D_h(u,z).$$

Proof idea of Theorem 13.3. Prepare four named pieces and then combine them. *Primal progress* controls the minimized gradient-model residual at the query point x. *Dual progress* controls the z-part and uses strong convexity of h once to produce a norm-square residual. *Coupling relation via convexity* decomposes the gap at x into a y-part and a z-residual. The *transport inequality* introduces the proof-only point $v = (1-\tau)y + \tau z^+$ and upper bounds the primal residual by the transported dual residual. The algebraic rearrangement is postponed until the final combination.

Proof of Theorem 13.3. **Primal progress.** Define the minimized gradient-model residual

$$R_g := \min_{w \in X} \left\{ \langle g, w-x \rangle + \frac{L}{2} \|w-x\|^2 \right\}.$$

Since y^+ is a minimizer, Lemma 13.1 gives

$$f(y^+) \leq f(x) + R_g. \tag{A}$$

Dual progress. Apply Lemma 13.2 to the mirror function $(L/\rho)h$ and the scalar $1/\tau$. It first gives

$$\frac{1}{\tau}\left\langle g, z^+ - z \right\rangle + \frac{L}{\rho} D_h(z^+, z) \leq \frac{1}{\tau} \langle g, u-z \rangle + \frac{L}{\rho} D_h(u,z) - \frac{L}{\rho} D_h(u,z^+).$$

Since h is ρ-strongly convex, $(L/\rho)D_h(z^+,z) \geq (L/2)\left\|z^+ - z\right\|^2$. Therefore

$$\frac{1}{\tau}\left\langle g, z^+ - z \right\rangle + \frac{L}{2}\left\|z^+ - z\right\|^2 \leq \frac{1}{\tau} \langle g, u-z \rangle + \frac{L}{\rho}(D_h(u,z) - D_h(u,z^+)). \tag{B}$$

Coupling relation via convexity. By convexity at x,

$$f(x) - f(u) \leq \langle g, x-u \rangle = \langle g, x-z \rangle + \langle g, z-u \rangle.$$

Since $x - z = \frac{1-\tau}{\tau}(y-x)$ and convexity gives $\langle g, y-x \rangle \leq f(y) - f(x)$, multiplying by τ gives

$$\tau(f(x) - f(u)) \leq (1-\tau)(f(y) - f(x)) + \tau \langle g, z-u \rangle. \tag{C}$$

Transport inequality. Now introduce the proof-only point

$$v := (1-\tau)y + \tau z^+.$$

Since $x = (1-\tau)y + \tau z$, we have

$$v - x = \tau(z^+ - z), \qquad z - z^+ = \frac{x-v}{\tau}.$$

Recall that $R_g = \min_{w\in X}\{\langle g, w-x\rangle + (L/2)\,\|w-x\|^2\}$. Evaluating this model at v gives

$$R_g \le \langle g, v-x\rangle + \frac{L}{2}\,\|v-x\|^2 = \tau\left\langle g, z^+ - z\right\rangle + \frac{L\tau^2}{2}\left\|z^+ - z\right\|^2. \tag{D}$$

Linear coupling. Take the weighted sum

$$(A) + (D) + \tau^2\cdot(B) + (C).$$

This gives directly

$$\tau\big(f(x) - f(u)\big) + f(y^+) \le (1-\tau)\big(f(y) - f(x)\big) + \frac{L\tau^2}{\rho}\big(D_h(u,z) - D_h(u,z^+)\big) + f(x).$$

The $f(x)$-terms cancel after rearranging, and the result is

$$f(y^+) - f(u) + \frac{L\tau^2}{\rho}D_h(u,z^+) \le (1-\tau)\big(f(y) - f(u)\big) + \frac{L\tau^2}{\rho}D_h(u,z).$$

Multiplying by $\rho/(L\tau^2)$ gives

$$\frac{\rho}{L\tau^2}\big(f(y^+) - f(u)\big) + D_h(u,z^+) \le \frac{\rho(1-\tau)}{L\tau^2}\big(f(y) - f(u)\big) + D_h(u,z),$$

as claimed. □

Theorem 13.4 (Accelerated smooth convex rate)**.** *Assume that E, X, f, h, L, ρ satisfy the smoothness, convexity, and strong-convexity hypotheses in Theorem 13.3. Let $x^\star \in \arg\min_{x\in X} f(x)$. Choose $z_0 \in X$ and any $y_0 \in X$, and run Definition 13.1 with*

$$\tau_t := \frac{2}{t+2}, \qquad t = 0, 1, 2, \ldots.$$

Then, for every $T \ge 1$,

$$f(y_T) - f(x^\star) \le \frac{4L\,D_h(x^\star, z_0)}{\rho(T+1)^2}.$$

In particular, if $X = E = \mathbb{R}^n$, $h(x) = \frac{1}{2}\,\|x\|_2^2$, and $\rho = 1$, then

$$f(y_T) - f(x^\star) \le \frac{2L\,\|x^\star - z_0\|_2^2}{(T+1)^2}.$$

Proof of Theorem 13.4. Apply Theorem 13.3 at step t with $u = x^\star$. For $t \geq 1$, the choice $\tau_t = 2/(t+2)$ gives

$$\frac{1-\tau_t}{\tau_t^2} = \frac{t(t+2)}{4} \leq \frac{(t+1)^2}{4} = \frac{1}{\tau_{t-1}^2}.$$

For $t = 0$, $\tau_0 = 1$, so the coefficient $(1-\tau_0)/\tau_0^2$ is zero. Thus the potential

$$\frac{\rho}{L\tau_t^2}(f(y_{t+1}) - f(x^\star)) + D_h(x^\star, z_{t+1})$$

is at most $D_h(x^\star, z_0)$ after telescoping from $t = 0$ to $t = T-1$. Since $\tau_{T-1} = 2/(T+1)$, we get

$$\frac{\rho(T+1)^2}{4L}(f(y_T) - f(x^\star)) \leq D_h(x^\star, z_0).$$

This proves the first bound. For $h(x) = \frac{1}{2}\|x\|_2^2$, $D_h(x^\star, z_0) = \frac{1}{2}\|x^\star - z_0\|_2^2$. □

For smooth strongly convex objectives, the simplest way to get the accelerated linear rate is to restart the convex accelerated method. In a general norm, this requires one extra compatibility assumption: the Bregman radius at the beginning of each epoch must be upper bounded by the squared norm.

Theorem 13.5 (Restarted acceleration for smooth strongly convex objectives)**.** *Assume that E, X, f, h, L, ρ satisfy the hypotheses in Theorem 13.4. Assume moreover that f is μ-strongly convex with respect to $\|\cdot\|$, and that there exists $M > 0$ such that*

$$D_h(a, b) \leq \frac{M}{2}\|a - b\|^2 \qquad \forall a, b \in X.$$

Let $x^\star \in \arg\min_{x\in X} f(x)$. Fix an epoch length $N \geq 1$ and define $w_0 \in X$. For each epoch $k = 0, 1, 2, \ldots$, run Definition 13.1 for N steps with

$$y_0^{(k)} = z_0^{(k)} = w_k, \qquad \tau_t = \frac{2}{t+2}, \qquad t = 0, \ldots, N-1,$$

and set $w_{k+1} := y_N^{(k)}$. Then

$$f(w_k) - f(x^\star) \leq \left(\frac{4LM}{\rho\mu(N+1)^2}\right)^k (f(w_0) - f(x^\star)).$$

In particular, if $N + 1 \geq \sqrt{8LM/(\rho\mu)}$, then

$$f(w_k) - f(x^\star) \leq 2^{-k}(f(w_0) - f(x^\star)).$$

Thus, with $N + 1 \asymp \sqrt{LM/(\rho\mu)}$, reaching $f(w_k) - f(x^\star) \leq \epsilon$ takes

$$O\left(\sqrt{\frac{LM}{\rho\mu}}\log\frac{f(w_0) - f(x^\star)}{\epsilon}\right)$$

gradient evaluations. In the Euclidean special case $X = E = \mathbb{R}^n$, $h(x) = \frac{1}{2}\|x\|_2^2$, and $\rho = M = 1$, this becomes

$$O\left(\sqrt{\frac{L}{\mu}}\log\frac{f(w_0) - f(x^\star)}{\epsilon}\right).$$

Proof of Theorem 13.5. Apply Theorem 13.4 to epoch k, with $z_0^{(k)} = w_k$. This gives

$$f(w_{k+1}) - f(x^\star) \leq \frac{4L}{\rho(N+1)^2} D_h(x^\star, w_k).$$

Using the upper bound on D_h,

$$f(w_{k+1}) - f(x^\star) \leq \frac{2LM}{\rho(N+1)^2} \left\| x^\star - w_k \right\|^2.$$

By μ-strong convexity,

$$f(w_k) - f(x^\star) \geq \frac{\mu}{2} \left\| w_k - x^\star \right\|^2.$$

Combining the two inequalities gives

$$f(w_{k+1}) - f(x^\star) \leq \frac{4LM}{\rho\mu(N+1)^2} (f(w_k) - f(x^\star)).$$

Iterating proves the first claim, and the displayed choice of N makes the contraction factor at most $1/2$. □

Euclidean Nesterov equivalence. In the unconstrained Euclidean case $X = E = \mathbb{R}^n$ and $h(w) = \frac{1}{2} \left\| w \right\|_2^2$, linear coupling is exactly the usual two-sequence Nesterov accelerated gradient method [Nes04], written in different variables. One standard form of NAG is

$$\begin{aligned} \tilde{x}_{t+1} &= \tilde{y}_{t+1} + \beta_t(\tilde{y}_{t+1} - \tilde{y}_t), \\ \tilde{y}_{t+2} &= \tilde{x}_{t+1} - \frac{1}{L} \nabla f(\tilde{x}_{t+1}), \end{aligned}$$

where $\beta_t = (s_t - 1)/s_{t+1}$ and the coefficients satisfy $s_{t+1}^2 - s_{t+1} \leq s_t^2$. For the simple choice $s_t = (t+2)/2$, this gives $\beta_t = t/(t+3)$. Now set $s_t = 1/\tau_t$ in the Euclidean linear-coupling iteration. The two updates are

$$y_{t+1} = x_t - \frac{1}{L} \nabla f(x_t), \qquad z_{t+1} = z_t - \frac{1}{L\tau_t} \nabla f(x_t).$$

Since $x_t - y_{t+1} = \frac{1}{L} \nabla f(x_t)$ and $x_t = (1-\tau_t) y_t + \tau_t z_t$,

$$z_{t+1} = z_t - \frac{x_t - y_{t+1}}{\tau_t} = y_t + \frac{1}{\tau_t}(y_{t+1} - y_t).$$

Therefore

$$\begin{aligned} x_{t+1} &= (1 - \tau_{t+1}) y_{t+1} + \tau_{t+1} z_{t+1} \\ &= y_{t+1} + \frac{\tau_{t+1}(1-\tau_t)}{\tau_t}(y_{t+1} - y_t) \\ &= y_{t+1} + \frac{s_t - 1}{s_{t+1}}(y_{t+1} - y_t). \end{aligned}$$

Thus, after identifying $\tilde{x}_{t+1} = x_{t+1}$ and $\tilde{y}_{t+1} = y_{t+1}$, the Euclidean linear-coupling scheme is precisely NAG.

Dependency and proof sketch

1. Lemma 13.1 is the primal-progress inequality from the projected gradient model.
2. Lemma 13.2 is the dual-progress inequality from Lecture 8, rewritten so the last two terms form the residual that will be paired with smoothness.
3. Theorem 13.3 is the single technical inequality of the lecture. It couples primal progress at x with dual progress at z.
4. Theorem 13.4 follows by choosing $\tau_t = 2/(t+2)$, using $(1-\tau_t)/\tau_t^2 \leq 1/\tau_{t-1}^2$, and telescoping.
5. Theorem 13.5 restarts the convex accelerated guarantee and uses strong convexity to convert the starting distance of each epoch into the starting objective gap.

Exercises

1. **Bonus.** Formulate and prove a non-restarted linear-coupling theorem for smooth strongly convex objectives. Your answer should specify the algorithm, the parameter choice, the assumptions, the potential, and the accelerated linear rate.

References

[AZO17] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, volume 67 of *Leibniz International Proceedings in Informatics*, pages 3:1–3:22. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.