Lecture 13: Linear Coupling and Acceleration

TTIC 31070 / CMSC 35470 / BUSF 36903 / STAT 31015

Convex Optimization

Prof. Zhiyuan Li

Spring 2026

From the Lower Bound to the Algorithm

Lecture 12 (lower bound). For smooth L-smooth convex optimization in radius R:

$$\inf_{A} \sup_{f} \left[f(x_T) - f^\star \right] \;\geq\; \Omega\left(\frac{LR^2}{T^2} \right).$$

Where we are. Gradient descent gives $O(LR^2/T)$ — off by a factor of T.

Today. Construct an algorithm achieving $O(LR^2/T^2)$ — matching the lower bound.

Viewpoint: Linear coupling (Allen-Zhu & Orecchia, "Linear coupling: an ultimate unification of gradient and mirror descent," ITCS 2017). One gradient step gives primal progress, one mirror step gives dual progress, and a *coupled query point* makes the two terms match.

Color convention used today. y sequence (gradient step), z sequence (mirror step), x sequence (queries).

Euclidean Preview: Two Uses of the Same Gradient

Setup (intuition only, Euclidean). Query at x, set $g := \nabla f(x)$.

Use 1: Gradient step.

$$y^+ := x - \tfrac{1}{L} g.$$

Smoothness gives

$$f(y^+) \;\le\; f(x) - \tfrac{1}{2L}\|g\|_2^2.$$

Large gradient $\Rightarrow$ instant objective decrease.

Use 2: Mirror step.

$$z^+ := z - \alpha g.$$

For any comparator u,

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

Small gradient $\Rightarrow$ linear signal still becomes telescoping distance decrease.

Two regimes, one gradient. Gradient step likes $\|g\|$ *big*; mirror step is fine when $\|g\|$ is *small*. **Acceleration couples them so no case split is needed.**

Primal Progress

Lemma 13.1 (Primal progress). f is L-smooth on X; $x \in X$, $g = \nabla f(x)$, and

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

Then for **every** $v \in X$:

$$\boxed{f(y^+) \;\le\; f(x) + \langle g, v - x \rangle + \tfrac{L}{2}\|v - x\|^2.}$$

We will eventually pick v to be the auxiliary point that makes the mirror displacement match a gradient-model displacement.

Dual Progress

Lemma 13.2 (Dual progress). X convex; h convex and differentiable on a neighborhood of X; $z \in X$, $g \in E^*$, $\alpha > 0$, and

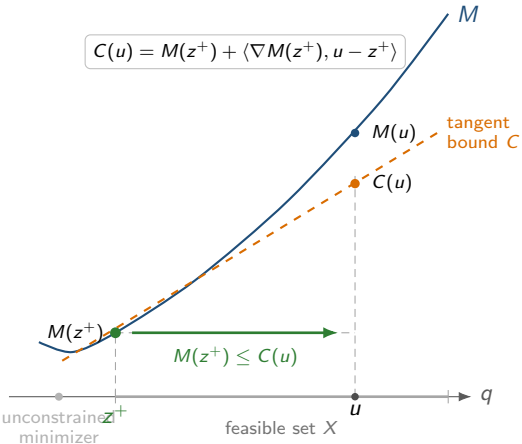
$$z^+ \in \operatorname*{argmin}_{q \in X} \Big\{ \alpha \langle g, q - z \rangle + D_h(q, z) \Big\}.$$

Then for **every** $u \in X$:

$$\boxed{\alpha \langle g, z^+ - z \rangle + D_h(z^+, z) \;\le\; \alpha \langle g, u - z \rangle + D_h(u, z) - D_h(u, z^+).}$$

The pair $D_h(u, z) - D_h(u, z^+)$ is the **telescoping potential drop**.

Three-Point Inequality: Tangent-Bound View



Recall the three-point inequality for

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

The mirror step chooses $z^+ \in \operatorname{argmin}_{q \in X} M(q)$.

At a constrained minimizer, $\nabla M(z^+)$ need not vanish. The first-order condition gives

$$\langle \nabla M(z^+), u - z^+\rangle \ge 0, \quad u \in X.$$

Thus

$$M(z^+) \le C(u) := M(z^+) + \langle \nabla M(z^+), u - z^+\rangle.$$

Expanding this tangent value gives

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

which is the three-point inequality used in Lemma 13.2.

Linear Coupling: The Update

Zeyuan Allen-Zhu

Lorenzo Orecchia

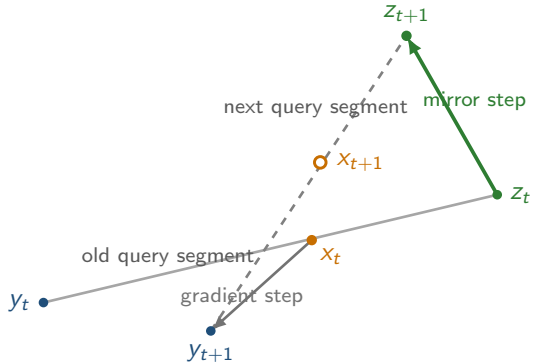
Definition 13.1 (Linear coupling step). Given $y_t, z_t \in X$ and $\tau_t \in (0,1]$:

- **Coupled query.** $x_t := (1-\tau_t)y_t + \tau_t z_t, \quad g_t := \nabla f(x_t)$.
- **Gradient step (at x_t).** $y_{t+1} = \operatorname{argmin}_{q\in X}\left\{\langle g_t, q - x_t\rangle + \frac{L}{2}\|q - x_t\|^2\right\}$.
- **Mirror step (at z_t).** $z_{t+1} = \operatorname{argmin}_{q\in X}\left\{\frac{1}{\tau_t}\langle g_t, q - z_t\rangle + \frac{L}{\rho}D_h(q, z_t)\right\}$.

Three sequences, one gradient per step.

- x_t: where we evaluate ∇f — a *convex combination of y_t and z_t*.
- y_t: tracks gradient-step descent; the *output sequence*.
- z_t: tracks mirror-step distance; the *dual-progress sequence*.

Linear Coupling: Geometry of the Iterates



$$x_t = (1-\tau_t)y_t + \tau_t z_t, \qquad x_{t+1} = (1-\tau_{t+1})y_{t+1} + \tau_{t+1} z_{t+1}.$$

- y_{t+1} comes from the gradient step at x_t.
- z_{t+1} comes from the mirror step at z_t.
- x_{t+1} is recomputed only after both updates, using the next weight τ_{t+1}.

One-Step Coupling Inequality

Theorem 13.3 (One-step linear-coupling). f, h differentiable; f convex and L-smooth; h ρ-strongly convex. With y, z, τ, x, y^+, z^+ as in Definition 13.1, for every $u \in X$:

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

Let $\Phi(y, z; \tau) := \frac{\rho}{L\tau^2}(f(y) - f(u)) + D_h(u, z)$. The inequality says the new state's potential at τ is dominated by the old state's potential at τ, but with the function-gap weight *shrunk by* $(1-\tau)$.

Why this is acceleration.

- GD-style descent gives f-decrease at y^+ weighted by the *unshrunk* potential.
- Coupling produces an extra $(1-\tau)$ factor on the *old* f-gap — exactly the slack a telescoping sum needs.

Proof Piece A: Primal Progress

Define the minimized gradient-model residual at the query point x:

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

Since y^+ minimizes the same model, Lemma 13.1 gives

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

Role of A. This is the only place where smoothness enters. The gradient step at x contributes the best possible quadratic-model residual.

Proof Piece B: Dual Progress

Apply Lemma 13.2 to the mirror step in Definition 13.1, i.e. to the mirror function $\tilde{h} = (L/\rho)h$ and scalar $\alpha = 1/\tau$. Since $D_{\tilde{h}} = (L/\rho)D_h$,

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

Multiplying by τ^2 gives a dual-progress bound at the scale needed for the coupling proof. Since h is ρ-strongly convex,

$$\frac{L\tau^2}{\rho}D_h(z^+, z) \ge \frac{L\tau^2}{2}\|z^+ - z\|^2.$$

Therefore, with $R_z := \tau\langle g, z^+ - z\rangle + \frac{L\tau^2}{2}\|z^+ - z\|^2$,

$$\boxed{R_z \le \tau\langle g, u - z\rangle + \frac{L\tau^2}{\rho}\big(D_h(u, z) - D_h(u, z^+)\big).} \qquad \text{(B)}$$

Proof Piece C: Coupling via Convexity

Convexity at the query point x gives

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

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

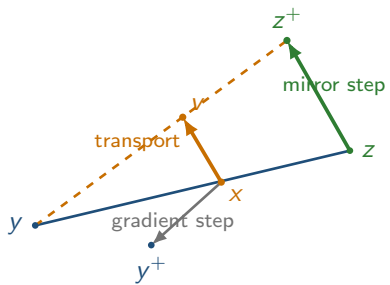
$$x - z = \frac{1 - \tau}{\tau}(y - x).$$

Convexity again gives $\langle g, y - x \rangle \le f(y) - f(x)$. Therefore

$$\boxed{\tau\big(f(x) - f(u)\big) \le (1 - \tau)\big(f(y) - f(x)\big) + \tau\langle g, z - u \rangle.} \qquad \text{(C)}$$

Role of C. It turns the comparator gap at x into an old y-gap plus the linear term that will cancel dual progress.

Proof Piece D: Transport Inequality



Proof-only point. Use the same convex-combination weight as x:

$$x = (1-\tau)y + \tau z, \qquad v = (1-\tau)y + \tau z^+.$$

It is not an iterate. Recall

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

Since $v - x = \tau(z^+ - z)$, the minimized primal residual is no larger than the transported dual residual:

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

Role of D. A mirror displacement at z gives a feasible gradient-model displacement at x. This is the transport step in the proof.

Proof Recipe: Combine the Pieces

Take the weighted sum

$$\boxed{(A) + (D) + (B) + (C).}$$

The two linear leftovers cancel:

$$\tau\langle g, u - z\rangle + \tau\langle g, z - u\rangle = 0, \qquad R_g \le R_z.$$

What remains 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 Theorem 13.3:

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

Telescoping: Choosing τ_t

Potential. Pick $u = x^\star$. Define

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

One-step (Theorem 13.3).

$$\Phi_t \;\le\; \frac{\rho(1-\tau_t)}{L\tau_t^2}\big(f(y_t) - f^\star\big) + D_h(x^\star, z_t).$$

Telescope condition. Need $\dfrac{1-\tau_t}{\tau_t^2} \le \dfrac{1}{\tau_{t-1}^2}$ so the right side is dominated by Φ_{t-1}.

Choice $\tau_t = \dfrac{2}{t+2}$**.** Then

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

For $t = 0$: $\tau_0 = 1 \Rightarrow$ coefficient is 0, so the chain starts cleanly.

Accelerated Smooth Convex Rate

Theorem 13.4 (Accelerated rate). f convex L-smooth; h ρ-strongly convex. With $\tau_t = 2/(t+2)$ and $y_0, z_0 \in X$:

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

Euclidean specialization. $X = \mathbb{R}^n$, $h(z) = \frac{1}{2}\|z\|_2^2$, $\rho = 1$:

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

This matches Lecture 12's lower bound $\Omega(LR^2/T^2)$ up to constants. ✓

Proof sketch. Telescope $\Phi_t \le \Phi_{t-1}$ from $t = 0$ to $T-1$, then divide out $\rho/(L\tau_{T-1}^2) = \rho(T+1)^2/(4L)$ from the function-gap weight at T. Initial $\Phi_{-1} = D_h(x^\star, z_0)$ since $\tau_0 = 1$ makes the convex-combination weight zero.

Smooth Strongly Convex: Restart

Idea. Run the convex accelerated method for N steps, restart with new $z_0 = y_N$.

Each epoch (use Theorem 13.4 + extra Bregman bound $D_h(a,b) \le \frac{M}{2}\|a-b\|^2$):

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

Strong convexity of f converts distance to gap: $\|x^\star - w_k\|^2 \le \frac{2}{\mu}(f(w_k) - f^\star)$. Hence

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

Choose $N+1 \ge \sqrt{8LM/(\rho\mu)}$**:** contraction $\le 1/2$ per epoch.

Smooth Strongly Convex: Accelerated Linear Rate

Theorem 13.5 (Restarted acceleration). f convex L-smooth and μ-strongly convex; h ρ-strongly convex with $D_h(a,b) \le \frac{M}{2}\|a-b\|^2$. With $N+1 \asymp \sqrt{LM/(\rho\mu)}$ steps per epoch and K epochs:

$$\boxed{f(w_K) - f^\star \;\le\; 2^{-K}\big(f(w_0) - f^\star\big).}$$

Total complexity: $O\big(\sqrt{LM/(\rho\mu)}\,\log(1/\varepsilon)\big)$ gradient evaluations.
In Euclidean ($\rho = M = 1$): $O\big(\sqrt{L/\mu}\,\log(1/\varepsilon)\big)$.

Compare with non-accelerated.

- GD (smooth strongly convex): $O(\kappa \log(1/\varepsilon))$, $\kappa = L/\mu$.
- Restarted AGD: $O(\sqrt{\kappa}\log(1/\varepsilon))$. **Improvement: $\sqrt{\kappa}$.**

This matches Lecture 12's lower bound $\Omega(\sqrt{\kappa}\log(1/\varepsilon))$. ✓

Why Acceleration Works (Recap)

In one sentence. Same gradient at x_t feeds two updates that produce *cancellable residuals*.

- **Gradient step at** x_t gives primal progress through a quadratic upper model at any v.
- **Mirror step at** z_t gives dual progress through a Bregman potential drop at any u.
- **Coupled query** $x_t = (1 - \tau_t)y_t + \tau_t z_t$ + transport identity: the mirror residual at z_t *equals* the gradient-model residual at x_t when v is chosen as the auxiliary $v_t = (1 - \tau_t)y_t + \tau_t z_{t+1}$.

What changes vs. GD. Not the gradient information, not the smoothness, not the strong convexity. **Only how the same gradient is used twice**: primal progress + dual progress.

Cost: one gradient per step (same as GD), three sequences to maintain (vs. one).

Euclidean Specialization: Nesterov Momentum Form

Yu. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," Soviet Math. Dokl. 27 (1983).

Setting. $X = \mathbb{R}^n$, $h(z) = \frac{1}{2}\|z\|_2^2$, $\rho = 1$.

Linear coupling becomes: $y_{t+1} = x_t - \frac{1}{L} g_t, \quad z_{t+1} = z_t - \frac{1}{L\tau_t} g_t.$

Algebra: $z_{t+1} = y_t + \frac{1}{\tau_t}(y_{t+1} - y_t)$, so

$$x_{t+1} = (1 - \tau_{t+1}) y_{t+1} + \tau_{t+1} z_{t+1} \;=\; y_{t+1} + \beta_t (y_{t+1} - y_t),$$

with $\beta_t = \tau_{t+1}(1 - \tau_t)/\tau_t = (s_t - 1)/s_{t+1}$, $s_t = 1/\tau_t$.

Nesterov's Accelerated Gradient (1983).

$$\tilde{x}_{t+1} = \tilde{y}_{t+1} + \beta_t (\tilde{y}_{t+1} - \tilde{y}_t),$$

$$\tilde{y}_{t+2} = \tilde{x}_{t+1} - \frac{1}{L} \nabla f(\tilde{x}_{t+1}).$$

Yurii Nesterov

Same algorithm, different variables. Linear coupling $\Leftrightarrow$ NAG via $(x, y, z) \leftrightarrow (\tilde{x}, \tilde{y})$ with z implicit.

Bonus Problem: Non-Restart Strongly Convex Coupling

Task. Give a theorem statement and proof for a *non-restarted* linear-coupling method for smooth strongly convex objectives.

Your answer should specify:

- the algorithm and parameter choice;
- the assumptions, including the geometry used by the mirror step;
- the potential that contracts;
- the resulting accelerated linear rate.

Bonus policy. This problem is worth at most 10 bonus points. You may use LLMs, but you must check the statement and proof carefully. The worst possible outcome is -3 bonus points, and a correct proof is guaranteed not to receive negative bonus points.

Summary

Linear coupling. Three sequences y_t, z_t, x_t, one gradient per step.

- Gradient step at x_t (descent).
- Mirror step at z_t (telescoping potential).
- Coupled query $x_t = (1-\tau_t)y_t + \tau_t z_t$ + transport identity $\Rightarrow$ residuals cancel.

Rates.

- Smooth convex: $f(y_T) - f^\star \le 4LD_h(x^\star, z_0)/(\rho(T+1)^2)$. $O(LR^2/T^2)$ — **tight.**
- Smooth strongly convex (restart): $O(\sqrt{\kappa}\log(1/\varepsilon))$. $\sqrt{\kappa}$ **vs** κ **for GD.**

Lecture 12 + Lecture 13 = tight. Acceleration is exactly the algorithm that closes the smooth Euclidean lower bound.

Next. Second-order methods: Hessians, local quadratic models, and Newton's method.

Bibliographic Notes

- **Yu. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," Soviet Math. Dokl. 27 (1983).** Original accelerated gradient method via estimating sequences.
- **Yu. Nesterov, Introductory Lectures on Convex Optimization, Kluwer (2004).** Textbook treatment; constant-step and adaptive variants for smooth convex / smooth strongly convex.
- **Z. Allen-Zhu & L. Orecchia, "Linear coupling: an ultimate unification of gradient and mirror descent," ITCS (2017).** Linear coupling viewpoint: the coupled query + primal/dual progress derivation used today.
- **S. Bubeck, Convex Optimization: Algorithms and Complexity, FnT (2015).** Modern survey; Section 3.7 gives a compact linear-coupling treatment.