AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

March 15, 2026

Outline

Midterm Review

Sample Midterm Q&A

Convexity

Convexity: $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all x, y,

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle \,.$$

Jensen's inequality: for any $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall\, x, y,$$

A useful generalization is $f(\frac{1}{n}\sum_{i=1}^n x_i) \leq \frac{1}{n}\sum_{i=1}^n f(x_i)$.

Strong Convexity:

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \,.$$

Properties of strong convexity:

- Unique minimizer x^* exists.
- Quadratic growth: $f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2$.
- Gradient is strongly monotone:
  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$.

Smoothness

Definition: f is L-smooth if ∇f is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \;\leq\; L\,\|x - y\|\,.$$

Descent Lemma (Quadratic Upper Bound):

$$f(y) \;\leq\; f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\,\|y - x\|^2\,.$$

Refined Gradient Inequality:

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \tfrac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Co-coercivity: $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\,\|\nabla f(x) - \nabla f(y)\|^2$.

For convex L-smooth f:

$$\|\nabla f(x)\|^2 \;\leq\; 2L(f(x) - f^*).$$

Proof: minimize the RHS of descent lemma over y to obtain $f^* \leq \min_y$ RHS; the result follows by rearranging terms.

Lipschitzness

Definition: f is L-Lipschitz if

$$|f(x) - f(y)| \leq L \|x - y\|, \quad \forall\, x, y.$$

Subgradient Bound: If f is convex and L-Lipschitz, then $\|v\| \leq L$ for all $v \in \partial f(x)$.

Proof:

- For $v \in \partial f(x)$: $f(x + v) - f(x) \geq \langle v, v \rangle = \|v\|^2$.
- Lipschitz gives $|f(x + v) - f(x)| \leq L \|v\|$.
- Combine: $\|v\|^2 \leq L \|v\| \Rightarrow \|v\| \leq L$.

What about f is convex and continuous on $\mathbb{R}^n$? Can you show f Lipschitz on any compact set?

Gradient Descent

Update Rule:

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

Three Regimes with Different Proof Strategies:

- L-smooth (nonconvex): sufficient decrease $\rightarrow$ telescope on f.
- Convex + smooth: distance decrease $\|x_k - x^*\|^2 \rightarrow$ telescope.
- SC + smooth: contraction $\|x_{k+1} - x^*\|^2 \leq q \|x_k - x^*\|^2$.

Common paradigm: find a one-step progress inequality $\rightarrow$ telescope $\rightarrow$ extract rate.

GD: Smooth Nonconvex Case

Setup: f is ℓ-smooth, bounded below by f_*, step size $\eta \leq 1/\ell$.

Sufficient Decrease: Apply the descent lemma at $y = x_{k+1}$, $x = x_k$:

$$f(x_{k+1}) \;\leq\; f(x_k) - \frac{\eta}{2} \left\|\nabla f(x_k)\right\|^2 .$$

Telescope $k = 0, \ldots, T-1$**:**

$$\frac{\eta}{2} \sum_{k=0}^{T-1} \left\|\nabla f(x_k)\right\|^2 \;\leq\; f(x_0) - f_*.$$

Extract Rate: $\min$ is bounded by the average, which is bounded by the telescoped sum.

Takeaway: Descent lemma + telescoping is the foundational technique.

GD: Convex Smooth

Setup: f convex and ℓ-smooth, step size $\eta \leq 1/\ell$.

One-step Relation:

$$2\eta\big(f(x_{k+1}) - f(x^*)\big) \;\leq\; \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 .$$

How to derive it:

- Expand $\|x_{k+1} - x^*\|^2$ and substitute $x_{k+1} = x_k - \eta \nabla f(x_k)$.
- Use convexity to bound $\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*$.
- Use co-coercivity or descent lemma to handle $\|\nabla f(x_k)\|^2$.

Then: Telescope + monotonicity $f(x_{k+1}) \leq f(x_k) \Rightarrow$ convergence of $f(x_T) - f^*$.

GD: Strongly Convex Smooth

Setup: f is α-SC and ℓ-smooth, $\eta = 1/\ell$.

Goal: Show a contraction in $\|x_{k+1} - x^*\|^2$.

Proof: Let $T(x) = x - \eta \nabla f(x)$, $d = x - y$, $g = \nabla f(x) - \nabla f(y)$.

- Expand: $\|T(x) - T(y)\|^2 = \|d\|^2 - 2\eta \langle g, d \rangle + \eta^2 \|g\|^2$.
- Co-coercivity: $\|g\|^2 \leq \ell \langle g, d \rangle \Rightarrow$ eliminate $\|g\|^2$.
- Strong monotonicity: $\langle g, d \rangle \geq \alpha \|d\|^2 \Rightarrow$ contraction factor $(1 - \alpha/\ell)$.

Takeaway: Co-coercivity + strong monotonicity together give linear convergence. The condition number $\kappa = \ell/\alpha$ controls the speed.

A Useful Trick: Minimizing a Quadratic Bound

Technique: Given an inequality involving a quadratic in some free variable, minimize over that variable to obtain the tightest possible bound.

Example: Sufficient Decrease from the Descent Lemma:

- Start: $f(x - \eta \nabla f(x)) \leq f(x) - \eta \left\|\nabla f(x)\right\|^2 + \frac{L\eta^2}{2} \left\|\nabla f(x)\right\|^2$.
- The RHS is a quadratic in η. Minimize it over η.
- Optimal $\eta^* = 1/L$ by setting derivative to zero.
- Plug back in: $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \left\|\nabla f(x_k)\right\|^2$.

Where this appears?

- Deriving the best step size.
- Converting between $\left\|\nabla f\right\|^2$ and $f(x) - f^*$ via quadratic lower bounds.

Subgradients and Subdifferential

Subgradient: $g \in \mathbb{R}^n$ is a subgradient of convex f at x if

$$f(y) \;\geq\; f(x) + \langle g, y - x \rangle\,, \quad \forall\, y.$$

Subdifferential: $\partial f(x) = \{\text{all subgradients at } x\}$.

Properties:

- $\partial f(x)$ is closed, convex, and nonempty for proper convex f at interior points.
- If f is differentiable at x: $\partial f(x) = \{\nabla f(x)\}$.
- Optimality: x^* minimizes f iff $0 \in \partial f(x^*)$.

Example: $f(x) = |x|$, $\partial f(0) = [-1, 1]$.

Subdifferential Calculus

Rules:

- Scaling: $\partial(af) = a\,\partial f$ for $a > 0$.
- Sum: $\partial(f+g) = \partial f + \partial g$ (under constraint qualification).
- Composition: $\partial g(x) = A^\top \partial f(Ax)$ for $g(x) := f(Ax)$.
- Pointwise max: $f(x) = \max_i f_i(x)$, then

$$\partial f(x) = \operatorname{conv}\left(\bigcup_{i:\, f_i(x)=f(x)} \partial f_i(x) \right).$$

Examples:

- Indicator ι_C: $\partial \iota_C(x) = N_C(x) = \{g : \langle g, y-x\rangle \le 0,\ \forall y \in C\}$.
- ℓ_1-norm: $v_i = \operatorname{sign}(x_i)$ if $x_i \neq 0$; $v_i \in [-1,1]$ if $x_i = 0$.

Subgradient Method

Update:

$$x_{k+1} \in x_k - \alpha_k \, g_k, \quad g_k \in \partial f(x_k).$$

Proof: (same as GD)

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle + \alpha_k^2 \|g_k\|^2 .$$

Then use:

- Convexity: $\langle g_k, x_k - x^* \rangle \geq f(x_k) - f^*$ to get a function-value decrease.
- Lipschitz bound: $\|g_k\| \leq L$ to control the error term.
- Telescope over k to accumulate function value progress.

Step sizes: $\sum \alpha_k = \infty$ (enough progress) and $\sum \alpha_k^2 < \infty$ (errors vanish) $\Rightarrow$ convergence.

Subgradient Method

Telescoped Bound:

$$2\sum_{i=1}^{k}\alpha_i\left(f(x_i)-f(x^*)\right) \;\leq\; R^2+\sum_{i=1}^{k}\alpha_i^2\left\|g_i\right\|^2, \quad R=\left\|x_0-x^*\right\|.$$

Constant Step Size $\alpha_k=\alpha$:

$$\min_{i=1,\ldots,k} f(x_i)-f^* \;\leq\; \frac{R^2}{2\alpha k}+\frac{\alpha L^2}{2} \quad\longrightarrow\quad \frac{\alpha L^2}{2} \text{ as } k\to\infty.$$

Converges to a neighborhood only; does not reach f^* exactly.

Polyak Step Size $\alpha_k=\frac{f(x_k)-f^*}{\|g_k\|^2}$:

$$\sum_{i=1}^{k}\frac{\left(f(x_i)-f^*\right)^2}{\left\|g_i\right\|^2} \;\leq\; R^2 \quad\Longrightarrow\quad \sum_{i=1}^{k}\left(f(x_i)-f^*\right)^2 \;\leq\; L^2R^2.$$

Projection

Projection: $\Pi_C(x) = \arg\min_{y \in C} \|x - y\|$, where C is closed convex.

Nonexpansiveness: (1-Lipschitz)

$$\|\Pi_C(x) - \Pi_C(y)\| \ \leq \ \|x - y\| \, .$$

Why It Matters:

- Projection never increases distances.
- For constrained problems, if you can bound $\|z_k - x^*\|$ before projection, the bound survives after projection.
- This is used whenever you analyze projected gradient or subgradient methods.

You should be familiar with other properties such as variational characterization, firmly nonexpansiveness and monotonicity.

Bregman Divergence

Definition: Given strictly convex differentiable $R : X \to \mathbb{R}$,

$$D_R(x, y) \;:=\; R(x) - R(y) - \langle \nabla R(y), x - y \rangle \,.$$

Properties:

- $D_R(x, y) \geq 0$ with equality iff $x = y$.
- Convex in x, but not necessarily convex in y.
- If R is σ-SC: $D_R(x, y) \geq \frac{\sigma}{2} \|x - y\|^2$.

Three-point Identity:

$$D_R(x, y) + D_R(y, z) \;=\; D_R(x, z) + \langle \nabla R(z) - \nabla R(y), x - y \rangle \,.$$

Examples:

- $R(x) = \frac{1}{2} \|x\|_2^2 \Rightarrow D_R = \frac{1}{2} \|x - y\|^2$.
- $R(x) = \sum x_i \log x_i \Rightarrow D_R = D_{\mathrm{KL}}$.

Mirror Descent

Primal Update:

$$x_{k+1} = \arg\min_{x \in X} \Big\{ D_R(x, x_k) + \alpha \langle g_k, x \rangle \Big\}, \quad g_k \in \partial f(x_k).$$

Equivalent Dual Space View:

$$\nabla R(x_{k+1}) = \nabla R(x_k) - \alpha_k \, g_k.$$

Interpretation:

- Map to dual: $y_k = \nabla R(x_k)$.
- Gradient step in dual: $y_{k+1} = y_k - \alpha_k g_k$.
- Map back: $x'_{k+1} = (\nabla R)^{-1}(y_{k+1})$.
- Project: $x_{k+1} = \arg\min_{z \in X} D_R(z, x'_{k+1})$.

Different choices of R adapt to different problem geometries.

- $R = \frac{1}{2} \|\cdot\|^2$ recovers projected subgradient;
- $R = \sum x_i \log x_i$ on the simplex recovers exponentiated gradient.

Mirror Descent

One-step Progress:

$$\alpha \langle g_k, x_k - x^* \rangle \;\leq\; D_R(x^*, x_k) - D_R(x^*, x_{k+1}) + \frac{\alpha^2}{2\sigma} \|g_k\|_*^2 .$$

Proof:

- Apply the three-point identity to relate $D_R(x^*, x_k)$, $D_R(x^*, x_{k+1})$, and $D_R(x_k, x_{k+1})$.
- Generalized Pythagorean theorem handles the projection step.
- Bound $D_R(x_k, x_{k+1}) \leq \frac{\alpha^2}{2\sigma} \|g_k\|_*^2$ using σ-strong convexity of R.

Then: Use convexity + telescope, exactly as in GD proofs. Only need to change ℓ_2-norm to Bregman divergence.

Stochastic Gradient Descent (SGD)

Objective:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

SGD Update:

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k), \quad i_k \sim \text{Unif}\{1, \ldots, n\} \text{ i.i.d.}$$

Key Properties:

- Unbiasedness: $\mathbb{E}_i[\nabla f_i(x)] = \nabla f(x)$.
- Variance decomposition:
  $\mathbb{E}_i \left\|\nabla f_i(x)\right\|^2 = \left\|\nabla f(x)\right\|^2 + \mathbb{E}_i \left\|\nabla f_i(x) - \nabla f(x)\right\|^2$.
- Bounded variance: If $\mathbb{E}_i \left\|\nabla f_i(x) - \nabla f(x)\right\|^2 \leq \sigma^2$, then
  $\mathbb{E}_i \left\|\nabla f_i(x)\right\|^2 \leq \left\|\nabla f(x)\right\|^2 + \sigma^2$.

SGD: Convex Case

Expand the squared distance: (same as GD)

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k \langle \nabla f_{i_k}(x_k), x_k - x^* \rangle + \alpha_k^2 \|\nabla f_{i_k}(x_k)\|^2 .$$

Take conditional expectation:

- Unbiasedness: $\mathbb{E}[\langle \nabla f_{i_k}, x_k - x^* \rangle \mid x_k] = \langle \nabla f(x_k), x_k - x^* \rangle$.
- Variance bound: $\mathbb{E}[\|\nabla f_{i_k}\|^2 \mid x_k] \leq \|\nabla f(x_k)\|^2 + \sigma^2$.
- Convexity: $\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*$.

Absorb the gradient norm: Use $\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*)$ and choose $\alpha \leq \frac{1}{2L}$ so the gradient term is absorbed into the LHS.

Telescope + Jensen: same as deterministic case, with an extra $\alpha^2\sigma^2$ per step from variance.

SGD: Strongly Convex Case

Extra Ingredient: After the same expansion + expectation,
$\mathbb{E}\left\|x_{k+1}-x^*\right\|^2 \leq (1-\alpha\mu)\left\|x_k-x^*\right\|^2 - 2\alpha(f(x_k)-f^*) + \alpha^2\mathbb{E}\left\|\nabla f_i(x_k)\right\|^2$.

Key Moves:

- Use a sharper second-moment bound that can vanish as iterates get closer to the minimizer to control the noise.
- Choose $\alpha \leq 1/(2L)$ so the $f(x_k) - f^*$ terms have the right sign and can be dropped.
- Result: $\mathbb{E}\left\|x_{k+1}-x^*\right\|^2 \leq (1-\alpha\mu)\mathbb{E}\left\|x_k-x^*\right\|^2 +$ noise.

Unroll the recursion: Apply the contraction repeatedly and sum the geometric series for the noise.

Takeaway: The linear convergence from SC is preserved, but with a noise proportional to $\alpha\sigma^2/\mu$ that does not vanish with constant step size.

Proof Techniques Summary

1. Expand–Bound–Telescope: Expand $\|x_{k+1} - x^*\|^2$; bound inner products using convexity or SC; telescope sums.

2. Descent Lemma: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$. Use it to show sufficient decrease.

3. Minimize a Quadratic Bound.

4. Variance Decomposition.

5. Bregman Three-point Identity: Replaces Euclidean distance expansion in mirror descent proofs.

6. Nonexpansiveness of Projection.

7. Refined Gradient Inequality.

Outline

Sample Q1(a)

Setup: (x_k) is quasi-Fejér monotone w.r.t. X if $\exists\, (\varepsilon_k) \geq 0$ summable s.t.

$$\|x_{k+1} - x\|^2 \;\leq\; \|x_k - x\|^2 + \varepsilon_k, \quad \forall\, x \in X,\ \forall\, k.$$

Part (a): Boundedness.

- **Technique:** Telescope the quasi-Fejér inequality.
- Fix any $x \in X$: $\|x_k - x\|^2 \leq \|x_0 - x\|^2 + \sum_{i=0}^{\infty} \varepsilon_i < \infty$.

Sample Q1(b)

Claim: If a subsequence $(x_{\varphi_k}) \to x^* \in X$, then $x_k \to x^*$.

Technique:

- Define tail sum $E_k = \sum_{i \geq k} \varepsilon_i \to 0$.
- For any $k' \geq \varphi_k$: telescope from φ_k to k':

$$\|x_{k'} - x^*\|^2 \;\leq\; \|x_{\varphi_k} - x^*\|^2 + E_{\varphi_k}.$$

- Both terms $\to 0$ as $k \to \infty$, so $\|x_{k'} - x^*\| \to 0$.

Sample Q1(c):

Claim: If f is convex and L-Lipschitz, then $\|v\| \leq L$ for all $v \in \partial f(x)$.

Technique:

- Subgradient inequality at direction $y = x + v$: $f(x+v) - f(x) \geq \|v\|^2$.
- Lipschitz: $|f(x+v) - f(x)| \leq L\,\|v\|$.
- Combine to get $\|v\| \leq L$.

Sample Q1(d)&(e)

Part (d): Quasi-Fejér Monotonicity Condition.

- Expand: $\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 + \alpha_k^2 \|v_k\|^2 - 2\alpha_k \langle v_k, x_k - x^* \rangle$.
- Convexity: $\langle v_k, x_k - x^* \rangle \geq f(x_k) - f^* \geq 0$.
- So $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \alpha_k^2 L^2$.
- **Condition:** $\sum \alpha_k^2 < \infty$ (so $\varepsilon_k = \alpha_k^2 L^2$ is summable).

Part (e): Convergence to Minimizer.

- From part (d), telescope to get $2 \sum \alpha_k (f(x_k) - f^*) \leq \|x_0 - x^*\|^2 + L^2 \sum \alpha_k^2$.
- If also $\sum \alpha_k = \infty$, then $\liminf f(x_k) = f^*$.
- Boundedness (part a) $\Rightarrow$ convergent subsequence to some $x^* \in X$.
- Part (b) $\Rightarrow$ full sequence converges to x^*.

Sample Q2(a)

Setup: f convex and L-smooth, g continuously differentiable and σ-SC.
Update: $\nabla g(x_{k+1}) = \nabla g(x_k) - \alpha_k \nabla f(x_k)$.

Part (a): Three-point Identity.

- **Technique:** Expand $D_g(x, y) + D_g(y, z)$ using the definition $D_g(a, b) = g(a) - g(b) - \langle \nabla g(b), a - b \rangle$.
- Algebraically regroup: $\langle \nabla g(z), x - z \rangle = \langle \nabla g(z), x - y \rangle + \langle \nabla g(z), y - z \rangle$.
- Result: $D_g(x, y) + D_g(y, z) = D_g(x, z) + \langle \nabla g(z) - \nabla g(y), x - y \rangle$.

Sample Q2(b)

Claim: $f(y) - f(x_k) \geq \frac{1}{\alpha_k}\big(D_g(y, x_{k+1}) - D_g(x_k, x_{k+1}) - D_g(y, x_k)\big)$.

Technique:

- Start from convexity: $f(y) - f(x_k) \geq \langle \nabla f(x_k), y - x_k \rangle$.
- Substitute $\nabla f(x_k) = \frac{1}{\alpha_k}(\nabla g(x_k) - \nabla g(x_{k+1}))$ from the update rule.
- Apply the three-point identity from part (a) to rewrite the inner product in terms of Bregman divergences.

Sample Q2(c)

Claim: With $\alpha_k = \sigma/L$, $D_g(x_k, x_{k+1}) \leq \alpha_k(f(x_k) - f(x_{k+1}))$.

Technique:

- Expand $D_g(x_k, x_{k+1})$ using the update rule to get $\alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle - D_g(x_{k+1}, x_k)$.
- Use L-smoothness of f: $\langle \nabla f(x_k), x_k - x_{k+1} \rangle \leq f(x_k) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2$.
- Use σ-SC of g: $D_g(x_{k+1}, x_k) \geq \frac{\sigma}{2} \|x_{k+1} - x_k\|^2$.
- Choose $\alpha_k = \sigma/L$ so the quadratic terms cancel exactly.

Sample Q2(d)

Claim: Show $O(1/T)$ convergence rate.

Technique:

- Set $y = x^*$ in part (b) and use part (c) to replace $D_g(x_k, x_{k+1})$:

$$f(x^*) - f(x_{k+1}) \;\geq\; \frac{1}{\alpha_k}\big(D_g(x^*, x_{k+1}) - D_g(x^*, x_k)\big).$$

- This is a telescoping inequality in $D_g(x^*, x_k)$.
- Sum $k = 0, \ldots, T-1$: the Bregman terms telescope, leaving $D_g(x^*, x_0)$.
- Since $\frac{1}{T}\sum f(x_k) \geq \min_k f(x_k)$, extract the convergence of $\min_k f(x_k) - f^*$.

Note: in fact, you can show last-iterate convergence rate $O(1/T)$ using monotonicity from part (c).

Sample Q3(a)

Setup: $f = \frac{1}{n}\sum f_i$, each f_i convex L-smooth, f is μ-SC.
$x_{k+1} = x_k - \alpha \nabla f_{i_k}(x_k)$, $i_k \sim \mathrm{Unif}[n]$.
Define $\sigma_f^* = \mathbb{E}\left\|\nabla f_i(x^*) - \nabla f(x^*)\right\|^2$.

Part (a): σ_f^* is independent of the choice of x^*.

- **Technique:** Use the refined gradient inequality (co-coercivity of individual f_i's).
- For any two minimizers x^*, x':
  $\frac{1}{2L}\mathbb{E}\left\|\nabla f_i(x') - \nabla f_i(x^*)\right\|^2 \le f(x') - f^* = 0.$
- So $\nabla f_i(x') = \nabla f_i(x^*)$ for all $i \Rightarrow \sigma_f^*$ is the same at any minimizer.

Sample Q3(b)

Claim: $\mathbb{E}\left\|\nabla f_i(x)\right\|^2 \leq 4L(f(x) - f^*) + 2\sigma_f^*$.

Technique:

- Split: $\left\|\nabla f_i(x)\right\|^2 \leq 2\left\|\nabla f_i(x) - \nabla f_i(x^*)\right\|^2 + 2\left\|\nabla f_i(x^*)\right\|^2$.
- First term: use the refined gradient inequality $\frac{1}{2L}\left\|\nabla f_i(x) - \nabla f_i(x^*)\right\|^2 \leq f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle$, take expectation, inner product vanishes.
- Second term: $\mathbb{E}\left\|\nabla f_i(x^*)\right\|^2 = \left\|\nabla f(x^*)\right\|^2 + \sigma_f^* = \sigma_f^*$ since $\nabla f(x^*) = 0$.

Sample Q3(c)

Technique:

- **Expand** $\|x_{k+1} - x^*\|^2$ and take conditional expectation (standard).
- Use strong convexity: $\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^* + \frac{\mu}{2} \|x_k - x^*\|^2$.
- Substitute the bound from part (b) for $\mathbb{E} \|\nabla f_i(x_k)\|^2$.
- Choose $\alpha \leq \frac{1}{2L}$ so the $f(x_k) - f^*$ terms can be dropped (they have the right sign).
- Arrive at: $\mathbb{E} \|x_{k+1} - x^*\|^2 \leq (1 - \alpha\mu)\mathbb{E} \|x_k - x^*\|^2 + 2\alpha^2\sigma_f^*$.

Unroll the recursion:

- Apply repeatedly: $(1 - \alpha\mu)^k \|x_0 - x^*\|^2 + 2\alpha^2\sigma_f^* \sum_{j=0}^{k-1}(1 - \alpha\mu)^j$.
- Geometric series: $\sum(1 - \alpha\mu)^j \leq \frac{1}{\alpha\mu}$.

Sample Q3(d)

Technique: From part (c), bound has two terms:

$$\underbrace{(1-\alpha\mu)^T \left\|x_0 - x^*\right\|^2}_{\text{optimization error}} + \underbrace{\frac{2\alpha\sigma_f^*}{\mu}}_{\text{noise}}.$$

Control each term separately:

- **Noise** $\leq \varepsilon/2$: requires $\alpha \leq \frac{\mu\varepsilon}{4\sigma_f^*}$.
- **Optimization error** $\leq \varepsilon/2$: requires $(1-\alpha\mu)^T \leq \frac{\varepsilon}{2\|x_0 - x^*\|^2}$.
- Use $\log(1-\alpha\mu) \leq -\alpha\mu$ to simplify: $T \geq \frac{1}{\alpha\mu} \log \frac{2\|x_0 - x^*\|^2}{\varepsilon}$.
- Take $\alpha = \min\{\frac{\mu\varepsilon}{4\sigma_f^*}, \frac{1}{2L}\}$ and the corresponding T.

Key Takeaways

Master the key inequalities: descent lemma, strong convexity, refined gradient inequality, co-coercivity, strong monotonicity.

The expand–bound–telescope paradigm appears in nearly every convergence proof.

Minimizing a quadratic bound is a useful trick (optimal step size, sufficient decrease, converting between quantities).

For SGD: variance decomposition + conditional expectation + absorbing gradient terms.

For mirror descent: Bregman divergence replaces Euclidean distance but the proof structure is the same.

Good Luck on the Midterm!