AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

February 4, 2026

Outline

Optimization Algorithm

Goal: solve

$$\min_{x \in \mathcal{X}} f(x)$$

An optimization algorithm is an iterative procedure:

$$x_0 \to x_1 \to x_2 \to \cdots$$

that uses limited information about f (and $\mathcal{X}$) to improve the objective.

Two core design questions:

- Update rule: how to compute x_{k+1} from x_k?
- Goal: where do you want x_k to move towards?

For ML purpose:

- Computation should be scalable, robust, and can stop anytime.
- Critical point? Local minima? Global minima?

Taxonomy

What information does the algorithm use at each step?

- **Zeroth-order:** only function values $f(x)$ (e.g., random search).
- **First-order:** gradients $\nabla f(x)$ or subgradients $\partial f(x)$.
- **Second-order:** Hessian $\nabla^2 f(x)$ or Hessian-vector products.

For large-scale ML:

- Zeroth-order is usually too sample-inefficient in high dimensions.
- Second-order is often too expensive per iteration.
- First-order methods dominate because gradients are informative *and* affordable (backprop/autodiff).

Today: the most basic first-order method: Gradient Descent (GD).

Gradient Descent (GD)

GD update:

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

Interpretation:

- $\nabla f(x_k)$ points in the direction of steepest increase locally.
- $-\nabla f(x_k)$ is the steepest descent direction under Euclidean geometry.
- $\eta_k > 0$ is called step size, which controls speed vs stability.

Why start with GD:

- It is the baseline behind SGD, momentum, Adam, PGD, etc.
- Its analysis is the simplest and well-established; provides insight for the analysis of other more advanced algorithms.

GD Example

Ridge regression:

$$f(w) = \frac{1}{2n}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2, \quad X \in \mathbb{R}^{n\times d},\ y \in \mathbb{R}^n.$$

Two ways to solve:

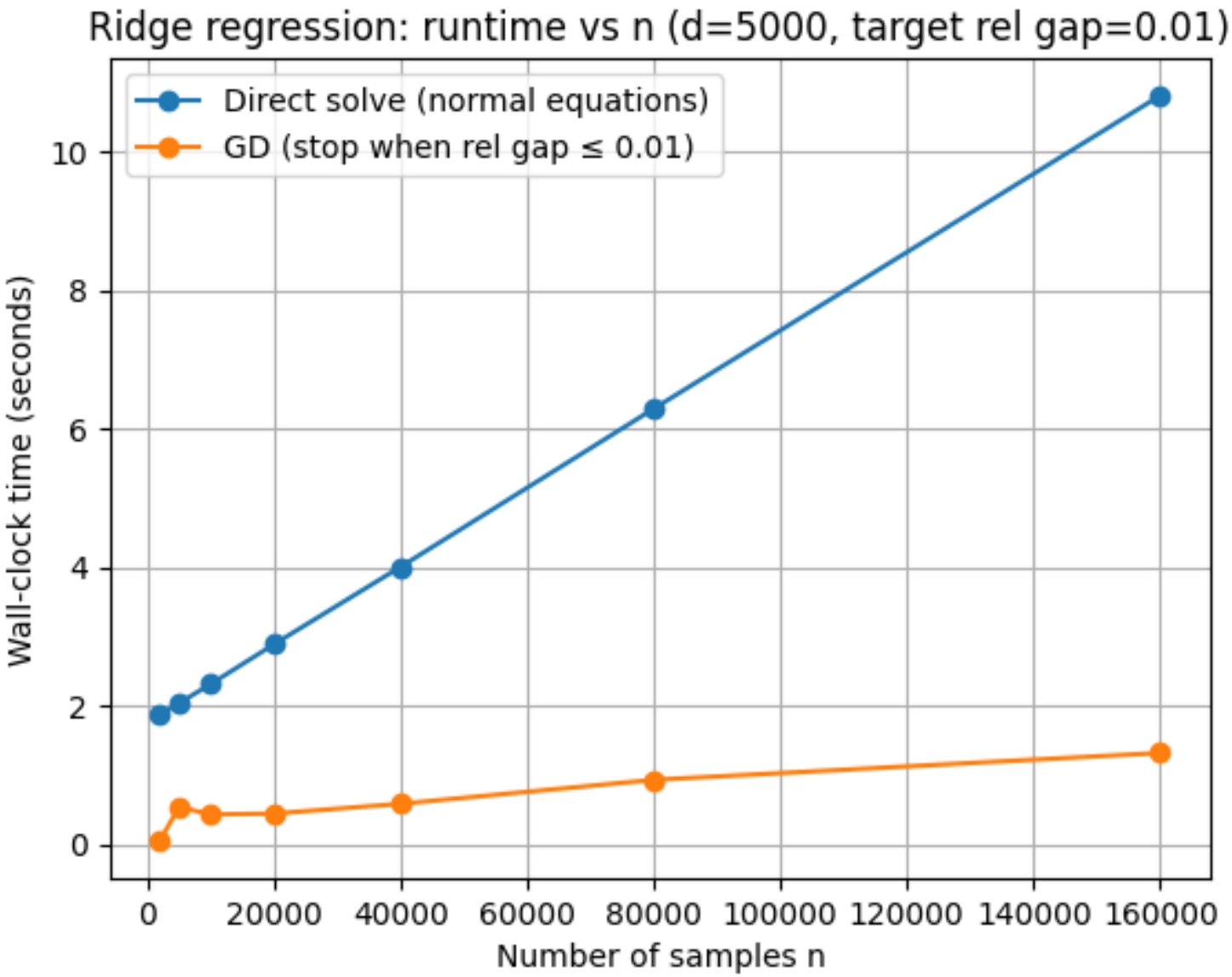
- Direct solve: compute the exact optimizer $w^\star$ by solving
$$\left(\frac{1}{n}X^\top X + \lambda I\right) w = \frac{1}{n}X^\top y.$$
- Gradient Descent: start from $w_0 = 0$ and iterate
$$w_{k+1} = w_k - \eta_k \nabla f(w_k), \quad \nabla f(w) = \frac{1}{n}X^\top(Xw - y) + \lambda w.$$

Stopping criterion:

$$\text{relgap}(w) = \frac{f(w) - f(w^\star)}{\max(|f(w^\star)|, \varepsilon)}.$$

Stop GD once $\text{relgap}(w_k) \le 0.01$.

Runtime Comparison



Outline

What Convergence Shall We Expect?

Assumption: f is ℓ-smooth and lower bounded.

GD with constant step sizes: $x_{k+1} = x_k - \eta \nabla f(x_k)$.

Goal: $\|\nabla f(x_k)\| \to 0$ as $k \to \infty$.

Convergence of GD for ℓ-smooth lower bounded function

Assume f is ℓ-smooth and bounded below by f_*. Run GD with step size $0 < \eta \leq 1/\ell$. Then for any $T \geq 1$, $\|\nabla f(x_T)\| \to 0$ as $T \to \infty$, and

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

Theory

Convergence of GD for ℓ-smooth lower bounded function

Assume f is ℓ-smooth and bounded below by f_*. Run GD with step size $0 < \eta \leq 1/\ell$. Then for any $T \geq 1$, $\|\nabla f(x_T)\| \to 0$ as $T \to \infty$, and

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

Proof sketch:

- Quadratic Taylor bound with $\eta \in (0, 1/\ell]$:
$$f(x_{k+1}) \leq f(x_k) - \tfrac{\eta}{2}\|\nabla f(x_k)\|^2.$$
- Sum from $k = 0$ to $T - 1$:
$$\tfrac{\eta}{2} \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq f(x_0) - f(x_T) \leq f(x_0) - f_*.$$
- Take minimum:
$$\min_{0 \leq k \leq T-1} \|\nabla f(x_k)\|^2 \leq \tfrac{1}{T} \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_*)}{\eta T}.$$

More Discussions

What if f is not lower bounded?

- E.g., $f(x) = -x^2$, then f diverges to $-\infty$.
- Not very interesting in ML.

What if f is not ℓ-smooth?

- E.g., $f(x) = x^4$?
- In ML, ℓ-smooth vs. non-ℓ-smooth, which is more common?

What if we use different step sizes?

- Larger constant step sizes, e.g., $\eta = 1.5/L$? $\eta = 2/L$? $\eta = 3/L$?
- Variable step sizes, e.g., η_k?
- Line-search?

Is the convergence rate optimal?

Outline

What convergence shall we expect?

Assumption: f is convex and L-Lipschitz (and differentiable).

GD with constant step sizes: $x_{k+1} = x_k - \eta \nabla f(x_k)$.

Goal: the global optimality gap $f(x_k) - f(x^\star) \to 0$ as $k \to \infty$.

Convergence of GD for convex L-Lipschitz function

Assume f is convex, L-Lipschitz, and differentiable. Let $x^\star \in \arg\min f$ and $\|x_1 - x^\star\| \le R$. Run GD for T steps with step size $\eta = \frac{R}{L\sqrt{T}}$ and define $\bar{x}_T := \frac{1}{T}\sum_{k=1}^{T} x_k$. Then

$$f(\bar{x}_T) - f(x^\star) \;\le\; \frac{RL}{\sqrt{T}}.$$

Theory

Theorem: Convergence of GD for convex L-Lipschitz function.

Proof sketch:

- Obtain relation between two consecutive steps:
  $\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - 2\eta\langle \nabla f(x_k), x_k - x^\star\rangle + \eta^2 \|\nabla f(x_k)\|^2.$
- Using convexity to eliminate variation by function value variation:
  $$\eta(f(x_k) - f(x^\star)) \leq \tfrac{1}{2}\|x_k - x^\star\|^2 - \tfrac{1}{2}\|x_{k+1} - x^\star\|^2 + \tfrac{\eta^2}{2}L^2.$$
- Sum $k = 1, \ldots, T$ and use $\|x_1 - x^\star\| \leq R$:
  $$\eta \sum_{k=1}^{T} (f(x_k) - f(x^\star)) \leq \tfrac{R^2}{2} + \tfrac{T\eta^2}{2}L^2.$$
- Use Jensen's inequality $f(\bar{x}_T) \leq \frac{1}{T}\sum_{k=1}^{T} f(x_k)$ to obtain
  $$f(\bar{x}_T) - f(x^\star) \leq \tfrac{R^2}{2\eta T} + \tfrac{\eta L^2}{2}.$$
- Choosing the best η to minimize the upper yields the desired result.

Smooth Convex

Assumption: f is convex and ℓ-smooth.

GD with constant step sizes: $x_{k+1} = x_k - \eta \nabla f(x_k)$.

Goal: the global optimality gap $f(x_k) - f(x^\star) \to 0$ as $k \to \infty$.

Convergence of GD for convex ℓ-smooth function

Assume f is convex and ℓ-smooth, and let $x^\star \in \arg\min f$. Run GD with step size $0 < \eta \leq 1/\ell$, then for any $T \geq 1$,

$$f(x_T) - f(x^\star) \;\leq\; \frac{\|x_0 - x^\star\|^2}{2\eta T}.$$

Theory

Convergence of GD for convex ℓ-smooth function

Assume f is convex and ℓ-smooth, and let $x^\star \in \arg\min f$. Run GD with step size $0 < \eta \leq 1/\ell$, then for any $T \geq 1$,

$$f(x_T) - f(x^\star) \;\leq\; \frac{\|x_0 - x^\star\|^2}{2\eta T}.$$

Proof sketch:

- Using convexity and ℓ-smoothness to show a stronger relation:
$$2\eta(f(x_{k+1}) - f(x^\star)) \leq \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2.$$
- Summing $k = 0, \ldots, T-1$ gives
$$2\eta \sum_{k=0}^{T-1} (f(x_{k+1}) - f(x^\star)) \leq \|x_0 - x^\star\|^2.$$
- Obtain monotonicity $f(x_k)$ from quadratic Taylor bound and
$$f(x_T) - f(x^\star) \leq \tfrac{1}{T} \sum_{k=0}^{T-1} (f(x_{k+1}) - f(x^\star)) \leq \frac{\|x_0 - x^\star\|^2}{2\eta T}.$$

More Discussions

Is there other convergence mode?

- What about $\|\nabla f(x_k)\|$?
- What about x_k?

Consider the Lyapunov function:

$$V(x_k) = f(x_k) - f(x^*) + \tfrac{L}{2}\|x_k - x^*\|^2.$$

What if there is no x^*?

- E.g., Logistic regression on linearly separable data [Ji and Telgarsky, 2020].

What if there is no ℓ-smooth?

- E.g., f is only C^2?

What if we use different step sizes?

- E.g., silver step sizes gives faster convergence [Altschuler and Parrilo, 2023].

Outline

What Convergence Shall We Expect?

Recall last lecture:

- Smooth: $\min_{k=1,\dots,T} \|\nabla f(x_k)\| = O(1/\sqrt{T})$.
- Lipschitz + convex: $f(\bar{x}_T) - f^\star = O(1/\sqrt{T})$.
- Smooth + convex: $f(x_T) - f^\star = O(1/T)$.

With strong convexity, we can do better:

- Lipschitz + strong convexity: $O(1/T)$.
- Smooth + strong convexity: **linear** convergence $O(e^{-T/\kappa})$.

Key parameter: condition number $\kappa := \ell/\alpha$ (smaller is better).

Strong Convexity

Definition: A differentiable function $f : \mathcal{X} \to \mathbb{R}$ is α-strongly convex if for all $x, y \in \mathcal{X}$,

$$f(y) \;\geq\; f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2.$$

Properties:

- There exists a unique minimizer for f.
- Quadratic growth: $f(x) - f(x^\star) \geq \frac{\alpha}{2}\|x - x^\star\|^2$.
- Gradient is strongly monotone:
  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2$.

Examples in ML:

- Least square problem where X has full column rank.
- Any convex loss with ℓ_2-regularizer.

Strongly Convex + Lipschitz

Assumption: f is α-strongly convex and L-Lipschitz (and differentiable).

GD with decreasing step sizes: $x_{k+1} = x_k - \eta_k \nabla f(x_k)$ with

$$\eta_k = \frac{2}{\alpha(k+1)}.$$

Goal: the global optimality gap $f(x_k) - f(x^\star) \to 0$ as $k \to \infty$.

Convergence of GD for strongly convex Lipschitz functions

Assume f is α-strongly convex, L-Lipschitz, and differentiable. Run GD with step sizes $\eta_k = \frac{2}{\alpha(k+1)}$ and define $\bar{x}_T := \frac{2}{T(T+1)} \sum_{k=1}^{T} k x_k$, then

$$f(\bar{x}_T) - f(x^\star) \;\leq\; \frac{2L^2}{\alpha(T+1)}.$$

Theory: Strongly Convex + Lipschitz

Theorem: Convergence of GD for α-strongly convex L-Lipschitz function.

Proof sketch:

- Expand the square:
  $\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - 2\eta_k\langle \nabla f(x_k), x_k - x^\star\rangle + \eta_k^2\|\nabla f(x_k)\|^2.$
- Use strong convexity to obtain
  $\|x_{k+1} - x^\star\|^2 \leq (1 - \eta_k\alpha)\|x_k - x^\star\|^2 - 2\eta_k(f(x_k) - f(x^\star)) + \eta_k^2 L^2.$
- Use L-Lipschitz and step sizes $\eta_k = \frac{2}{\alpha(k+1)}$,
  $f(x_k) - f(x^\star) \leq \frac{(k-1)\alpha}{4}\|x_k - x^\star\|^2 - \frac{(k+1)\alpha}{4}\|x_{k+1} - x^\star\|^2 + \frac{L^2}{\alpha(k+1)}.$
- Telescoping and applying (general) Jensen's inequality:
  $$f(\bar{x}_T) - f(x^\star) \leq \frac{2}{T(T+1)}\sum_{k=1}^{T} k(f(x_k) - f(x^\star)) \leq \frac{2L^2}{\alpha(T+1)}.$$

Strongly convex + Smooth

Assumption: f is α-strongly convex and ℓ-smooth.

GD with constant step size: $x_{k+1} = x_k - \eta \nabla f(x_k)$.

Goal: the global optimality gap $\|x_k - x^*\| \to 0$ as $k \to \infty$.

Convergence of GD for strongly convex smooth functions

Assume f is α-strongly convex and ℓ-smooth. Run GD with step size $\eta = \frac{1}{\ell}$. Then for all $t \geq 1$,

$$\|x_{t+1} - x^\star\|^2 \leq \left(1 - \frac{\alpha}{\ell}\right)^t \|x_1 - x^\star\|^2 \;\leq\; e^{-t\alpha/\ell} \|x_1 - x^\star\|^2.$$

Theory: strongly convex + smooth

Notation:

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

Proof sketch:

- Expand the difference after one GD step:
$$\|T(x) - T(y)\|^2 = \|d - \eta g\|^2 = \|d\|^2 - 2\eta \langle g, d \rangle + \eta^2 \|g\|^2.$$
- Co-coercivity: $\|g\|^2 \leq \ell \langle g, d \rangle$. With $\eta = 1/\ell$,
$$\|T(x) - T(y)\|^2 \leq \|d\|^2 - \frac{1}{\ell} \langle g, d \rangle.$$
- Strong monotonicity: $\langle g, d \rangle \geq \alpha \|d\|^2$. Thus
$$\|T(x) - T(y)\|^2 \leq \left(1 - \frac{\alpha}{\ell}\right) \|d\|^2.$$
- Take $y = x^\star$ to get
$$\|x_{k+1} - x^\star\|^2 \leq \left(1 - \frac{\alpha}{\ell}\right) \|x_k - x^\star\|^2 \leq e^{-k\alpha/\ell} \|x_1 - x^\star\|^2.$$

More Discussions

Is strong convexity necessary for linear convergence?

- No in terms of objective. The PL condition suffices:

$$\|\nabla f(x)\|^2 \;\geq\; 2\mu(f(x) - f^\star).$$

Condition number and speed:

- If κ is large (ill-conditioned), GD is slow. This motivates Conjugate Gradient (for quadratics) and acceleration (next week).

Is the GD rate $O(\kappa \log(1/\varepsilon))$ optimal?

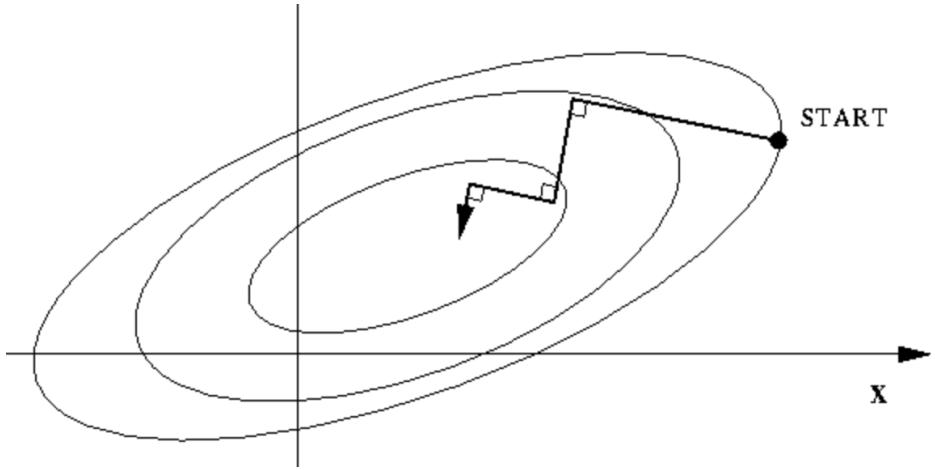
- GD with silver step sizes can obtain $O(\kappa^{0.7864} \log(1/\varepsilon))$ [Altschuler and Parrilo, 2025].
- Nesterov acceleration can obtain $O(\sqrt{\kappa} \log(1/\varepsilon))$ (next week).

Outline

Why do we need Conjugate Gradient (CG)?

Motivation: GD can be slow when the problem is ill-conditioned.

- For α-strongly convex, ℓ-smooth objectives, GD needs $O(\kappa \log(1/\varepsilon))$ iterations, where $\kappa = \ell/\alpha$.
- Large κ creates narrow valleys $\Rightarrow$ zig-zagging and slow progress.



CG is tailored for SPD quadratics:

- Minimizing $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ with symmetric $A \succ 0$;
- equivalently, solving the SPD linear system $Ax = b$.

Ill-conditioning in ML: correlated features, imbalanced data, etc.

SPD Quadratics

Problem: solve $Ax = b$ with $A \succ 0$.

Equivalent minimization:

$$f(x) := \frac{1}{2} x^\top A x - b^\top x. \qquad \nabla f(x) = Ax - b.$$

Let $x^\star = A^{-1} b$.

Residual and error:

$$r_k := b - Ax_k = -\nabla f(x_k), \qquad e_k := x_k - x^\star.$$

Key identity: with the notation $\|v\|_A := \sqrt{v^\top A v}$,

$$f(x) - f(x^\star) = \frac{1}{2}(x - x^\star)^\top A (x - x^\star) = \frac{1}{2}\|x - x^\star\|_A^2.$$

What goes wrong with GD?

GD with exact line search:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \text{ where } \alpha_k \in \arg\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k)).$$

A consequence of exact line search:

$$\langle \nabla f(x_{k+1}), \nabla f(x_k) \rangle = 0.$$

Why this still can be slow when $\kappa(A)$ is large:

- Level sets of f are elongated ellipses.
- The gradient is perpendicular to level sets $\Rightarrow$ steps bounce across the valley.
- Orthogonality above is in the *Euclidean* geometry, but the quadratic is governed by the *A-geometry*.

Takeaway: we need directions that respect the curvature induced by A.

Geometric idea: A-conjugate directions

Definition: $p, q \neq 0$ are A-conjugate if $p^\top A q = 0$.

Why A-conjugacy is the way to go?

- For quadratics, progress is measured in the A-norm:
$$f(x) - f(x^\star) = \tfrac{1}{2}\|x - x^\star\|_A^2.$$
- No undoing under exact line searches: if $x^+ = x + \alpha p$ then $\nabla f(x^+) = \nabla f(x) + \alpha A p$. So for any other direction q,
$$\langle \nabla f(x^+), q \rangle = \langle \nabla f(x), q \rangle + \alpha\, q^\top A p.$$
If $q^\top A p = 0$, the component along q is unchanged.
- Hence, once we minimize along a direction, later steps along A-conjugate directions do not re-introduce error in that direction.

Krylov Subspace

Krylov subspace generated by A and r_0:

$$\mathcal{K}_k(A, r_0) := \operatorname{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{k-1} r_0\}.$$

For SPD A, the CG iterate x_k satisfies

$$x_k \in \arg \min_{x \in x_0 + \mathcal{K}_k(A, r_0)} f(x).$$

Interpretation: each iteration adds one Krylov direction, and CG re-optimizes over the whole subspace.

CG updates

Initialize:

$$r_0 = b - Ax_0, \qquad p_0 = r_0.$$

For $k = 0, 1, 2, \ldots$**:**

$$\alpha_k = \frac{r_k^\top r_k}{p_k^\top A p_k}, \qquad x_{k+1} = x_k + \alpha_k p_k, \qquad r_{k+1} = r_k - \alpha_k A p_k,$$

$$\beta_{k+1} = \frac{r_{k+1}^\top r_{k+1}}{r_k^\top r_k}, \qquad p_{k+1} = r_{k+1} + \beta_{k+1} p_k.$$

Cost: one matrix-vector product and a few dot products per iteration.

Why this update direction?

Goal: pick directions that are A-conjugate and stay in the Krylov subspace.

Stay in Krylov subspace: $r_{k+1} \in \mathcal{K}_{k+2}(A, r_0)$ and $p_k \in \mathcal{K}_{k+1}(A, r_0) \subseteq \mathcal{K}_{k+2}(A, r_0)$, so p_{k+1} remains in $\mathcal{K}_{k+2}$.

Efficient recurrence: use only linear combination of two vectors:

$$p_{k+1} \in \operatorname{span}\{r_{k+1}, p_k\} \implies p_{k+1} = r_{k+1} + \beta_{k+1} p_k.$$

Choose β_{k+1} to enforce A-conjugacy: pick β_{k+1} so that $p_{k+1}^\top A p_k = 0$.

An Equivalent Update

Start from: $x_{k+1} = x_k + \alpha_k p_k$ and $p_k = r_k + \beta_k p_{k-1}$.

Since $p_{k-1} = \dfrac{x_k - x_{k-1}}{\alpha_{k-1}}$, we have

$$p_k = r_k + \beta_k \frac{x_k - x_{k-1}}{\alpha_{k-1}}.$$

Plug into $x_{k+1} = x_k + \alpha_k p_k$:

$$\boxed{x_{k+1} = x_k + \alpha_k r_k + \underbrace{\left(\beta_k \frac{\alpha_k}{\alpha_{k-1}}\right)}_{=:\,\gamma_k}(x_k - x_{k-1}).}$$

Looks like momentum, but: coefficients come from inner products and enforce Krylov optimality and A-conjugacy.

Properties

The iterates of CG satisfy:

- **Energy decreases:** $f(x_{k+1}) \leq f(x_k)$ (exact line search along p_k).
- **Residual orthogonality:** $r_i^\top r_j = 0$ for $i \neq j$.
- A**-conjugacy of directions:** $p_i^\top A p_j = 0$ for $i \neq j$.
- **Finite termination:** $x_k = x^\star$ in at most n steps (more precisely, degree of the minimal polynomial).

Practical note: in floating point arithmetic, conjugacy is approximate; CG typically converges fast but not exactly in n steps.

Convergence

Convergence of CG

Let $e_k = x_k - x^\star$. Then

$$\|e_k\|_A \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|e_0\|_A, \qquad \kappa = \kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Interpretation: to reach $\|e_k\|_A \le \varepsilon$, it suffices to take

$$k = O\Big(\sqrt{\kappa}\log(1/\varepsilon)\Big).$$

Compare: GD is $O(\kappa \log(1/\varepsilon))$ on the same quadratic.

Wrap-up and Next Week

Week 2: Key takeaways

- **Gradient Descent (GD):** $x_{k+1} = x_k - \eta \nabla f(x_k)$.
- **Nonconvex smooth:** constant step size $\eta \leq 1/L \Rightarrow O(1/\sqrt{T})$.
- **Convex:** Lipschitz $\Rightarrow O(1/\sqrt{T})$, and L-smooth $\Rightarrow O(1/T)$.
- **Strongly convex:** linear convergence; GD is slow when condition number κ large.
- **Conjugate Gradient (CG):** minimizing quadratics using A-conjugate directions, often faster than GD when κ is large.

Thank you for your attendance!

Questions?

Next week (Week 3): Momentum Methods and Acceleration

References


Altschuler, J. M. and Parrilo, P. A. (2023).
Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization.
arXiv preprint arXiv:2309.16530.


Altschuler, J. M. and Parrilo, P. A. (2025).
Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule.
Journal of the ACM, 72(2):1–38.


Ji, Z. and Telgarsky, M. (2020).
Directional convergence and alignment in deep learning.
Advances in Neural Information Processing Systems, 33:17176–17186.