Lecture 6

Last time
- Momentum
- Accelerated GD

Today
- Preconditioning
- Newton's method

Preconditioning

Last time we saw a way of accelerating GD doing some complicated gymnastics with three sequences. Today, we will see another way to speed up algorithms by simply changing the model we minimize at each iteration. The idea is simple, recall

$$\text{(GD)}\quad w_{k+1} \leftarrow \operatorname*{argmin}_z \left\{ \langle \nabla f(w_k), z - w_k \rangle + \frac{1}{2\eta_k} \|z - w_k\|^2 \right\}.$$

Why are we using the Euclidean norm to measure distances? For badly conditioned quadratics this yields very slow convergence!

Def (Mahalanobis norm): Given a symmetric positive definite matrix $Q \succ 0$, define the norm

$$\|w\|_Q = \sqrt{w^T Q w}.$$

Thus, given a sequence of matrices $P_1, P_2, \ldots$ setting the "right" geometry at iteration k, we consider

$$w_{k+1} \leftarrow \operatorname*{argmin}_z \left\{ \langle \nabla f(w_k), z - w_k \rangle + \frac{1}{2\eta_k} \|z - w_k\|_{P_k}^2 \right\} \quad \text{(PGD)}$$

$$= w_k - \eta_k P_k^{-1} \nabla f(w_k).$$

Recall that positive definite matrices might be written like

$$A = U D U^T$$

(D: Rescaling. U, U^T: Change of basis back and forth.)

Thus, preconditioning rescales certain directions via D^{-1}, hopefully making conditioning better; see demo.

There are several natural questions:

- Does preconditioning actually help?
- How to design good preconditioners?
- Are they efficient?

In the next few lectures we will study multiple preconditioners and give answers

to these questions. Let's start simple.

Fixed preconditioner throughout

Suppose we set we fixed $P_k = P$.
Then, the theory we study in Lectures 2 and 3 extend rather cleanly.

Def: We say that $f: \mathbb{R}^d \to \mathbb{R}$ is L_P-smooth with respect to $P \succ 0$ if

$$\|\nabla f(w_1) - \nabla f(w_2)\|_{P^{-1}} \le L_P \|w_1 - w_2\|_P \quad \forall w_1, w_2$$

The function f is μ_P-strongly convex wrt P if

$$\frac{\mu_P}{2}\|w_1 - w_2\|_P^2 + \langle \nabla f(w_1), w_2 - w_1 \rangle + f(w_1) \le f(w_2) \quad \forall w_1, w_2.$$

The function f satisfies μ_P-PL wrt P if

$$\mu_P (f(w) - f^*) \le \frac{1}{2}\|\nabla f(w)\|_{P^{-1}}^2. \quad \dashv$$

There is a very slick interpretation.

Lemma (Scale change): Let

$$\varphi(z) := f(P^{-1/2} z).$$

Suppose that $w_0, w_1, \ldots$ are the iterates of (PGD) applied to the function $f(\cdot)$. Consider (GD) applied to $\varphi(\cdot)$, i.e.,

$$z_{k+1} \leftarrow z_k - \eta_k \nabla \varphi(z_k),$$

with $z_0 = P^{1/2} w_0$. Then,

$$w_k = P^{-1/2} z_k \quad \text{and} \quad f(w_k) = \varphi(z_k).$$

This can be seen as commuting diagram

$$\begin{array}{ccc} z_0 & \xrightarrow{P^{-1/2}} & w_0 \\ \text{(GD) on } \varphi \downarrow & & \downarrow \text{(PGD) on } f. \\ z_k & \xleftarrow{P^{1/2}} & w_k \end{array}$$

Same results as before with different conditioning.

Using this result plus our results from Lecture 2 we easily derive.

Theorem (PGD convergence): Let f be L_P-smooth wrt $P \succ 0$, and run (PGD) with $\eta_k = 1/L_P$. Then, the following holds.

1. $$\min_{k \in [T-1]} \|\nabla f(w_k)\|_{P^{-1}}^2 \leq \frac{2 L_P (f(w_0) - f^*)}{T}.$$

2. Further, if f satisfies μ_P-PL wrt P,

$$f(w_k) - f^* \leq \left(1 - \frac{\mu_P}{L_P}\right)^k (f(w_0) - f^*) \quad \forall k.$$

Example: Consider a quadratic

$$f(x) = x^T A x + b^T x + c$$

with $A \succ 0$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}$. Then, a simple computation yields

$\mu_P = \lambda_{min}(P^{-1/2} A P^{-1/2})$ and $L_P = \lambda_{max}(P^{-1/2} A P^{-1/2})$.

Thus, (PGD) is faster than (GD) when

$$\kappa(A) > \kappa(P^{-1/2} A P^{-1/2}).$$

The best preconditioner would be $P = A$, but this corresponds to inverting a potentially arbitrary matrix, which has complexity $\Theta(d^3)$ (very expensive).

Newton method

Our original motivation for (PGD) was to change the geometry to get

faster complexity. But another motivation might be an even more accurate approximation. What's a good approximation?

$$w_{k+1} \leftarrow \arg\min_w \left\{ f(w_k) + \langle \nabla f(w_k), w - w_k \rangle + \frac{1}{2} \langle \nabla^2 f(w_k)(w - w_k), w - w_k \rangle \right\}.$$

Taylor's.

This step is not well defined since the approximation might not be convex when $\nabla^2 f(w_k)$ is no PD. Yet, we can use the optimality condition to derive

$$(N) \quad w_{k+1} \leftarrow w_k - \nabla^2 f(w_k)^{-1} \nabla f(w_k)$$

There are a few important comments:

▷ This method converges to critical points, not minimizers.

If we take $g(w) = -f(w)$, the method doesn't change.



▷ The method is affine invariant.

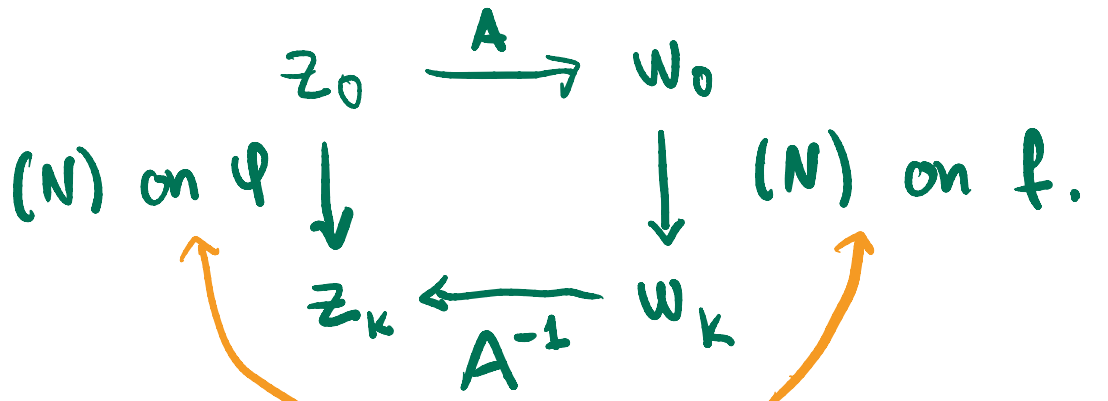
Lemma (affine invariance): Consider

$$\varphi(z) = f(Az)$$

with $A \in \mathbb{R}^{d \times d}$ invertible. Let $z_0, z_1, \ldots$ be the iterates of (N) applied to $\varphi(\cdot)$ and $w_0, w_1, \ldots$ be the iterates of (N) applied to $f(\cdot)$ with $z_0 = A^{-1} w_0$. Then,

$$z_k = A^{-1} w_k \qquad \forall k.$$

Again the commutative diagram looks like



Newton's doesn't care about change of variables.

This tells us that a good guarantee should not depend on the condition number.

Theorem (Quadratic convergence): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable with $w_\#$ s.t. $\nabla f(w_\#) = 0$. Suppose that there exists $r > 0$ s.t. $\forall w, u \in \underbrace{B_r(w_\#)}_{B}$,

(1) $\|\nabla^2 f(w) - \nabla^2 f(u)\|_{op} \le L_H \|w - u\|$;

(2) $\nabla^2 f(w) \succeq \mu I$; and

(3) $r < 2\mu/L_H$.

Then, for all $k \ge 0$,

$$\|w_{k+1} - w_\#\| \le \frac{L_H}{2\mu} \|w_k - w_\#\|^2, \quad \text{so}$$

$$\|w_{k+1} - w_\#\| \le \frac{2\mu}{L_H} \underbrace{\left(\frac{L_H}{2\mu} \|w_0 - w_\#\|\right)}_{<1}^{2^k}.$$

Proof: We use induction. Suppose $w_\# \in B$. Consider the map

$$t \longmapsto \nabla f(w_\# + t e_k), \qquad e_k = w_k - w_\#$$

Then, the chain rule gives that its Jacobian is given by

$$\nabla^2 f(\underbrace{w_\# + t e_k}_{\text{Stays in } B \text{ when } t \in [0,1].}) e_k.$$

By the fundamental theorem of calculus and $\nabla f(w_\#) = 0$,

$$(-) \quad \nabla f(w_k) = \int_0^1 \nabla^2 f(w_\# + t e_k) e_k \, dt.$$

Then

$$e_{k+1} = e_k - \nabla^2 f(w_k)^{-1} \nabla f(w_k)$$

$$= \nabla^2 f(w_k)^{-1} \int_0^1 \left(\nabla^2 f(w_k) - \nabla f^2(w_* + t e_k) \right) e_k \, dt$$

Taking norms and applying the triangle ineq, yields

$$\| e_{k+1} \| \leq \| \nabla f(w_k)^{-1} \|_{op} \int_0^1 \| \nabla^2 f(w_k) - \nabla^2 f(w_* + t e_k) \|_{op} \| e_k \| dt$$

(i) ↘ (ii) ↗ $\leq \frac{L_H}{\mu} \| e_k \|^2 \int_0^1 (1 - t) \, dt$

$$= \frac{L_H}{2\mu} \| e_k \|^2.$$

This implies $\frac{L_H}{2\mu} \| e_{k+1} \| \leq \frac{L_H^2}{4\mu^2} \| e_k \|^2$, unrolling this inequality using (iii) finishes the proof. □

In general, Newton might diverge, or cycle without ever converging (it generates beautiful fractals). This is also very expensive $O(d^3)$.