Lecture 11. Unconstrained optimization, gradient descent.

Zero-sum games. payoff matrix G. strategy $\in \mathbb{R}^n$.

Player X: for fixed x. player Y's best strategy is to minimize

$$\sum_{i,j} G_{ij} x_i y_j \Rightarrow X\text{'s goal is } \max_x \min_y \sum_{i,j} G_{ij} x_i y_j.$$

Player Y: for fixed y. player X's best strategy is to maximize

$$\sum_{i,j} G_{ij} x_i y_j \Rightarrow Y\text{'s goal is } \min_y \max_x \sum_{i,j} G_{ij} x_i y_j.$$

We claim they are dual problems. For example. $G = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$

If X choose strategy $x = (x_1, x_2)$. payoff of Y is $\begin{pmatrix} 3x_1 - 2x_2 \\ -x_1 + x_2 \end{pmatrix}$

goal of X: $\max_{x_1+x_2=1} \min \{ 3x_1 - 2x_2, -x_1 + 2x_2 \}$.

$\Leftrightarrow$ max z. s.t. $3x_1 - 2x_2 \geq z$, $-x_1 + x_2 \geq z$, $x_1 + x_2 = 1$, $x_1, x_2 \geq 0$.

If Y choose strategy $y = (y_1, y_2)$. payoff of X is $\begin{pmatrix} 3y_1 - y_2 \\ -2y_1 + y_2 \end{pmatrix}$.

goal of Y: $\min_{y_1+y_2=1} \max \{ 3y_1 - y_2, -2y_1 + y_2 \}$.

$\Leftrightarrow$ min w. s.t. $3y_1 - y_2 \leq w$, $-2y_1 + y_2 \leq w$, $y_1 + y_2 = 1$, $y_1, y_2 \geq 0$.

Minmax is the dual of maxmin. so equality holds by SD.

Theorem (von Neumann's Minimax Theorem).

$$\max_x \min_y x^T G y = \min_y \max_x x^T G y.$$

Theorem (Yao's min-max theorem).

$$\max_{x \in X} \mathbb{E}_A[c(A,x)] \geq \min_{a \in A} \mathbb{E}_X[c(a,X)].$$

Corollary (complementary slackness). Suppose x, y feasible for (P) (D).

Then x, y are optimal iff $y^T(b - Ax) = 0$, $x^T(A^T y - c) = 0$.

Proof. $c^T x \leq y^T A x \leq y^T b$. $\begin{cases} \text{either } y_i = 0, \text{ or } (Ax)_i = b_i \text{ tight} \\ \text{either } x_j = 0, \text{ or } (A^T y)_j = c_j \text{ tight} \end{cases}$ □

General convex optimization problems. (Unconstrained).

Consider an unconstrained smooth convex optimization $\min f(x)$.

Recall the first-order condition for optimality. $\nabla f(x) = 0$.

Example. $\min_w \|y - Xw\|^2$ $\quad w^* = (X^T X)^{-1} X^T y$.

However, if $f(x)$ becomes more complicated? $f(x) = x^2 + e^x$.

Since $f''(x) \geq 0$. $f'(x)$ increasing $\begin{cases} x = 0. \ f'(x) = 1 \Rightarrow x^* < 0. \\ x = -1. \ f'(x) = -2 + 1/e \Rightarrow x^* > -1 \end{cases}$

In general. f is convex then restriction to any line is also convex.

$\forall v \in \mathbb{R}^n$ $\begin{cases} \nabla f(x)^T v > 0 & x \to x - tv \\ \nabla f(x)^T v < 0 & x \to x + tv. \end{cases}$ for some $t > 0$.

On the other hand. we hope $f(x) > f(x + tv) \geq f(x) + t \nabla f(x)^T v$.

So v is a reasonable moving direction iff $\nabla f(x)^T v < 0$.

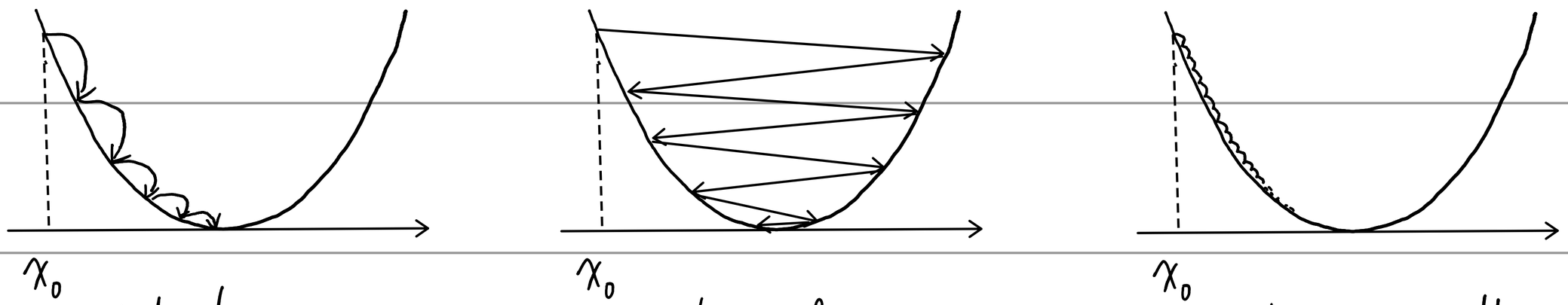
Descent method: $x_{k+1} = x_k + t_k v_k$ where $\nabla f(x_k)^T v_k < 0$.

Ideally stopping criterion : $\nabla f(x) = 0$. stopping at optimal.

Practical : $\|\nabla f(x)\| < \delta$. $|f(x_{k+1}) - f(x_k)| < \delta$ (or $\delta |f(x_k)|$).

or truncation if some fixed maximum # of steps is reached.

Two questions : 1. choose step size t_k ? 2. running time ?



condition $\nabla f(x)^T v < 0$ for direction, but no for t in general.

First attempt. simply select $v = -\nabla f(x)^T$. then $\nabla f(x)^T v < 0$ trivially.

Advantage : max rate descending direction. $\nabla_v f(x) = \nabla f(x)^T v$.

$-\|v\| \|\nabla f(x)\| \leq \nabla f(x)^T v \leq \|v\| \|\nabla f(x)\|$ by Cauchy-Schwarz.

with equality iff $v = \pm c \nabla f(x)$ for some $c > 0$.

Gradient descent : $x_{k+1} = x_k - t_k \nabla f(x_k)$. t_k is TBD.

Hope : $f(x_{k+1}) < f(x_k)$. consider $f(x) = ax^2$. $a > 0$. $x_{k+1} = x_k - 2atx_k$.

$f(x_{k+1}) < f(x_k) \Rightarrow (1-2at)^2 x_k^2 < x_k^2 \Rightarrow t < \frac{1}{a}$.

More generally. consider $f(x) = x^T Q x$. $Q \geq 0$. $x_{k+1} = x_k - 2tQx_k$.

$$f(x_{k+1}) = x_k^T Q x_k + 4t^2 (Qx_k)^T Q(Qx_k) - 4t (Qx_k)^T (Qx_k).$$

$f(x_{k+1}) < f(x_k)$ iff $t(Qx_k)^T Q(Qx_k) < (Qx_k)^T (Qx_k)$ $t < ?$

Proposition: $\lambda_{min} \|x\|_2^2 \le x^T Q x \le \lambda_{max} \|x\|_2^2$. (so $t < 1/\lambda_{max}$ suffice)

Proof. $Q \in \mathbb{R}^{n \times n}$ symmetric $\Rightarrow$ orthogonally diagonalize $Q = U \Lambda U^T$.

where $\Lambda = diag\{\lambda_1 \dots \lambda_n\}$ and $U^T U = I$. Assume $x = Uy$.

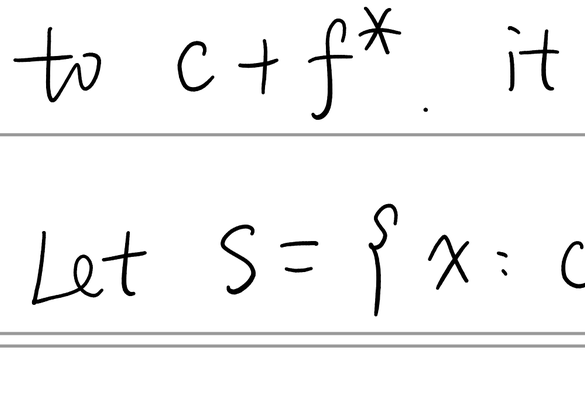
$$x^T Q x = y^T U^T Q U y = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 \le (\max \lambda_i) \|y\|_2^2.$$

$\|y\|_2^2 = y^T y = y^T U^T U y = x^T x = \|x\|_2^2$. Similar for λ_{min}. □

Now we know $f(x_{k+1}) < f(x_k)$. but does $f(x_k)$ converge to f^*?

Let $d_k = x_k - x^*$. and $g(v) = f(x^* + v) - f(x^*) = f(x^* + v) - f^*$

$g(d_k) = f(x_k) - f^*$. $g(d_{k+1}) < g(d_k)$ since $f(x_{k+1}) < f(x_k)$.



$g(d_k) \downarrow \ge 0$. $\Rightarrow \lim_{k \to \infty} g(d_k)$ exists $\triangleq c$

if $c = 0$ we're done. assume $c > 0$.

Intuitively that's impossible since $\forall x \in dom\, f$. $x_k = x$. we have $f(x_{k+1}) < f(x_k)$ as long as $f(x_k) \neq f^*$. if $f(x_k)$ sufficiently close to $c + f^*$. it is possible that $f(x_{k+1}) < c + f^* \Rightarrow g(d_{k+1}) < c$.

Let $S = \{x : c + f^* \le f(x) \le f(x_0)\}$ be a compact set.

$h(x) = f(x) - f(x - t\nabla f(x))$. h continuous if ∇f continuous.

$\Rightarrow \exists\, \delta = \min_{x \in S} h(x) > 0$. by our assumption.

$\Rightarrow f(x_{k+1}) < f(x_k) - \delta \iff g(d_{k+1}) < g(d_k) - \delta$. contradiction.

Lyapunov's global stability theorem in discrete time.

Suppose $d_{k+1} = p(d_k)$ where $p: \mathbb{R}^n \to \mathbb{R}^n$ continuous. $p(0) = 0$.

If there exists a (Lyapunov) function $l: \mathbb{R}^n \to \mathbb{R}$ continuous. and

① $l(0) = 0$, $l(x) > 0$ for all $x \neq 0$. (positivity).

② $l(x) \to \infty$ as $\|x\| \to \infty$. (radical unboundedness).

③ $l(p(x)) < l(x)$. $\forall x \neq 0$. (strict decrease).

Then $\forall d_0 \in \mathbb{R}^n$ we have $d_k \to 0$ as $k \to \infty$.

For gradient descent, set $p(v) = v - t\nabla f(x^* + v)$. $l(v) = g(v)$.

Question: ① how to choose t? ② how to know convergence rate?

In fact, if we can control t. then we can baund convergence rate.

A simple case: constant t. $\forall k \geq 0$. $t_k = t$ but may fail...

Example: $f(x) = |x|$. (approximate smoothly at 0). oscillation $-\frac{t}{2} \leftrightarrow \frac{t}{2}$.

Why? $f(x)$ changes rapidly near 0. Hope: $f(x)$ changes slowly.

Definition. (Lipschitz continuity) $|f(x)-f(y)| \le L|x-y|$.

In general. $f: \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous with $L>0$.

if $\|f(x)-f(y)\| \le L\|x-y\| \quad \forall x, y \in \mathbb{R}^n$. ($L$-Lipschitz)

Example. $f(x)=w^Tx$ is $\|w\|$-Lipschitz. (using l_2-norm)

$|w^T(x-y)| \le \|w\| \cdot \|x-y\|$ by Cauchy-Schwarz.

Example. $f(x)=Qx$ ($\mathbb{R}^n \to \mathbb{R}^n$) is $\lambda_{max}(Q)$-Lipschitz if $Q \succeq 0$.

$$\|f(x)-f(y)\| = \|Q(x-y)\| = \left((x-y)^T Q^2 (x-y)\right)^{1/2}$$

$$\le \left(\lambda_{max}(Q^2)\|x-y\|^2\right)^{1/2} = \lambda_{max}(Q)\|x-y\|.$$

Remark. If $Q \not\succeq 0$. then $\sigma_{max}(Q)$-Lipschitz (singular value).

If ∇f is L-Lipschitz. then $t<1/L$ is desired.