CS 331: Stochastic Gradient Descent Methods
Assignment 2

Lukang Sun, ID: 182056

September 14, 2021

p1. (1.)

$$\frac{||a||^2}{t} + t||b||^2 + 2\langle a, b\rangle = ||\frac{a}{\sqrt{t}} + \sqrt{t}b||^2 \geq 0,$$

this proves

$$\langle a, b\rangle \leq \frac{||a||^2}{2t} + \frac{t||b||^2}{2}. \tag{1}$$

(2.)

$$||a+b||^2 = ||a||^2 + ||b||^2 + 2\langle a, b\rangle,$$

since $2\langle a, b\rangle \leq ||a||^2 + ||b||^2$, we have finally

$$||a+b||^2 \leq 2||a||^2 + 2||b||^2 \tag{2}$$

(3.) from inequality(2), we know

$$||a||^2 = ||a+b-b||^2 \leq 2||a+b||^2 + 2||-b||^2 = 2||a+b||^2 + 2||b||^2,$$

move $2||b||^2$ to the left hand side, then both sides time 0.5, we have

$$\frac{1}{2}||a||^2 - ||b||^2 \leq ||a+b||^2 \tag{3}$$

p2. if $C = 0$, then we have

$$\mathbb{E}[||x^k - x^*||^2] \leq (1-\gamma\mu)^k ||x_0 - x^*||^2,$$

combine Markov inequality, for any $\varepsilon > 0$, we then have

$$\begin{aligned}\lim_{k\to\infty} \text{Prob}\left(\left\|X^k - X\right\| > \varepsilon\right) = \lim_{k\to\infty} \text{Prob}\left(\left\|X^k - X\right\|^2 > \varepsilon^2\right) &\leq \lim_{k\to\infty} \frac{\mathbb{E}[||x^k - x^*||^2]}{\varepsilon^2} \\ &\leq \lim_{k\to\infty} \frac{(1-\gamma\mu)^k||x_0 - x^*||^2}{\varepsilon^2} = 0.\end{aligned}$$

p3. denote $r^k = x^k - x^*$

$$\begin{aligned}\mathbb{E}[||r^{k+1}||^2 \mid x^k] &\le ||r^k||^2 + 2\gamma\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle + \gamma^2||\nabla f(x^k) - \nabla f(x^*)||^2 + \gamma^2\sigma^2 \\ &\le (1 - \frac{2\gamma\mu L}{\mu + L})||r^k||^2 - \gamma(\gamma - \frac{2}{\mu + L})||\nabla f(x^k) - \nabla f(x^*)||^2 + \gamma^2\sigma^2,\end{aligned} \tag{4}$$

where the second inequality uses

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \ge \frac{\mu L}{\mu + L}||x - y||^2 + \frac{1}{\mu + L}||\nabla f(x) - \nabla f(y)||^2.$$

We set $\gamma = \frac{2(1+\frac{\mu}{L})}{L(1+3\frac{\mu}{L})}$, then $\gamma \le \frac{2}{L}\frac{1}{1+\frac{2\frac{\mu}{L}}{1+\frac{\mu}{L}}} \le \frac{2}{L}\frac{1}{1+\frac{\mu}{L}} = \frac{2}{L+\mu}$, since $\frac{2}{1+\frac{\mu}{L}} \ge 1$. We insert this γ into (4), take whole expectation then we have

$$\mathbb{E}[||r^{k+1}||^2] \le (1 - \frac{4}{3 + \frac{L}{\mu}})\mathbb{E}[||r^k||^2] + \gamma^2\sigma^2, \tag{5}$$

use (5) iteratively for k from 0 to $i - 1$, then we have

$$\mathbb{E}[||r^i||^2] \le (1 - \frac{4}{3 + \frac{L}{\mu}})^i||r^0||^2 + \frac{\gamma^2\sigma^2}{\frac{2\gamma\mu L}{\mu+L}} \le (1 - \frac{4}{3 + \frac{L}{\mu}})^i||r^0||^2 + \frac{\gamma\sigma^2}{\mu},$$

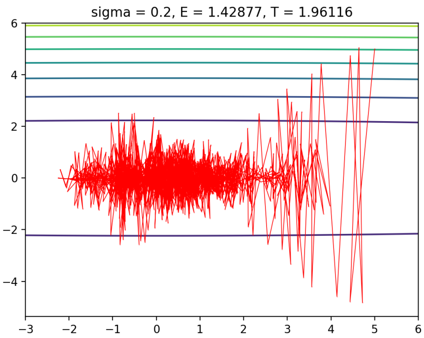
the second inequality uses $\frac{\mu+L}{2L} \le 1$. So in order to make $(1 - \frac{4}{3+\frac{L}{\mu}})^k \le \epsilon$, we only need

$$k \ge \frac{\frac{L}{\mu} + 3}{4}\log(\frac{1}{\epsilon}),$$

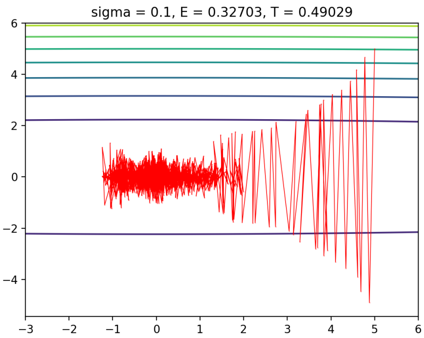
finally, we reach

$$\mathbb{E}[||x^k - x^*||^2] \le \epsilon||x^0 - x^*||^2 + \frac{\gamma\sigma^2}{\mu}.$$

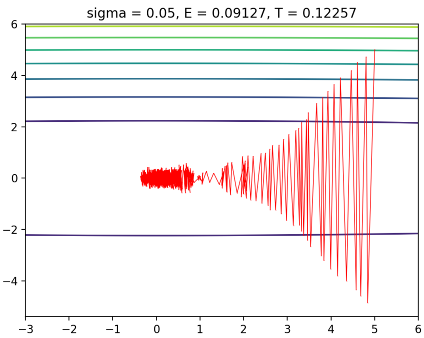
p4. In my first set of experiments(see Figure 1.), $A = \begin{bmatrix} 0.02 & 0 \\ 0 & 2 \end{bmatrix}, d = 2, x^0 = (5, 5), L = 2, \mu = 0.02, \gamma = \frac{2(1+\frac{\mu}{L})}{L(1+3\frac{\mu}{L})} \approx 0.9806, \epsilon = 10^{-10}, \frac{\frac{L}{\mu}+3}{4}\log(\frac{1}{\epsilon}) \approx 595, k = 1485$, I set σ equals 0.1, 0.2, 0.01, 0.05, 0.001, 0.005 separately, to estimate $\mathbb{E}\left[||x^k - x^*||^2\right]$, I sample x^k for 100 times and take the average value of $||x^k - x^*||^2$, I will denote the average value as E and $\frac{\gamma\sigma^2}{\mu}$ as T in the results. I use **numpy.random.normal()** to generate ξ.



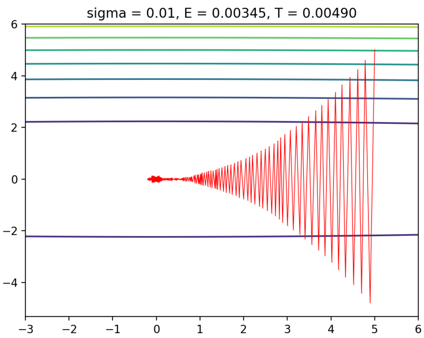
(a) $\gamma = 0.9806, \sigma = 0.2$



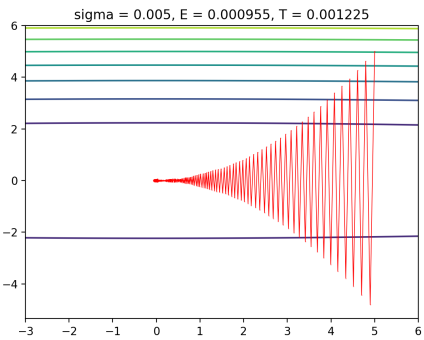
(b) $\gamma = 0.9806, \sigma = 0.1$



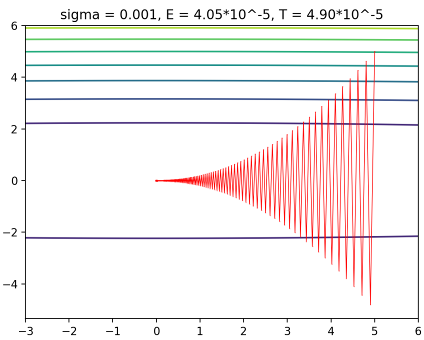
(c) $\gamma = 0.9806, \sigma = 0.05$



(d) $\gamma = 0.9806, \sigma = 0.01$

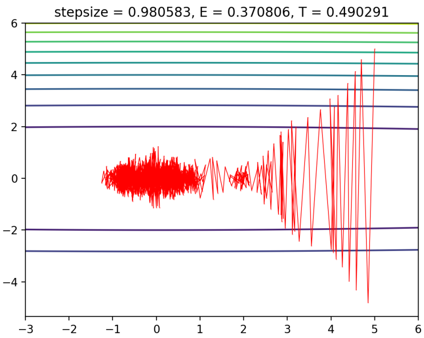


(e) $\gamma = 0.9806, \sigma = 0.005$

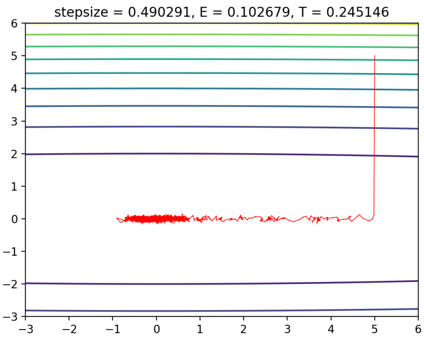


(f) $\gamma = 0.9806, \sigma = 0.001$

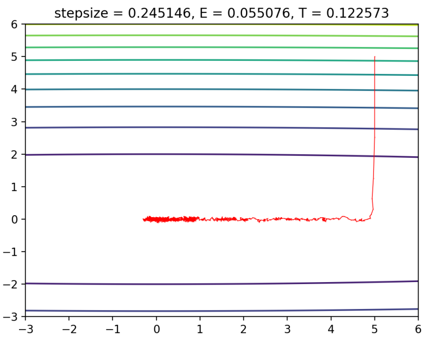
Figure 1: Experiments with different σ values, the horizontal axis is $x-axis$ and vertical axis is $y-axis$, the red trajectories show how SGD converges. From the results you can see that E and T always have same magnitude and have similar value, these testify that after enough iteration, SGD converges to a neighborhood of the optimal solution, it is hard for SGD to converge further after entering into the neighborhood and the neighborhood's size is well predicted by the theory(since T and E are similar).



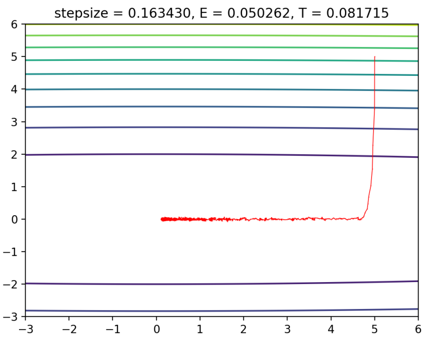
(a) $\gamma = 0.9806, \sigma = 0.1$



(b) $\gamma = 0.4903, \sigma = 0.1$



(c) $\gamma = 0.24515, \sigma = 0.1$



(d) $\gamma = 0.16343, \sigma = 0.1$

Figure 2: Experiments with different step sizes, these results testify that the size of the neighborhood has linear growth relationship with step size. From the results, you can also observe that $\frac{T}{E} \approx 2$, since from the derivation of problem 3., we know the theoretical neighborhood size is $\frac{\gamma^2\sigma^2}{\frac{2\gamma\mu L}{\mu+L}} \approx \frac{1}{2}\frac{\gamma\sigma^2}{\mu}$.

In my second set of experiments(see Figure 2.), $A = \begin{bmatrix} 0.02 & 0 \\ 0 & 2 \end{bmatrix}, d = 2, x^0 = (5,5), L = 2, \mu = 0.02, \sigma = 0.1, \epsilon = 10^{-10}, \frac{\frac{L}{\mu}+3}{4}\log(\frac{1}{\epsilon}) \approx 595, k = 1485$, I set step size γ approximately equals 0.9806, 0.4903, 0.24515, 0.16343 separately , to estimate $\mathbb{E}\left[||x^k - x^*||^2\right]$, I sample x^k for 100 times and take the average value of $||x^k - x^*||^2$, I will denote the average value as E and $\frac{\gamma\sigma^2}{\mu}$ as T in the results. I use **numpy.random.normal()**to generate ξ.