Foundations of Data Science and Machine Learning – *Homework 5*

Isaac Martin

Last compiled April 12, 2023

EXERCISE 1. Suppose $\mathbf{A}$ is a $n \times d$ full-rank matrix, with $n < d$, and fix $\mathbf{b} \in \mathbb{R}^n$. Consider minimizing the least squares objective $F(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ Note that in this setting, the solution space $\mathcal{S} = \{x : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ is an affine subspace of $\mathbb{R}^d$. We use gradient descent with constant step-size:

$$\mathbf{x} = \mathbf{x}_{k-1} - \eta \nabla F(\mathbf{x}_{k-1}).$$

(a) Give an upper bound for the step-size η such that gradient descent is guaranteed to converge for η below this threshold.

(b) Suppose that gradient descent is initialized at $\mathbf{x}_0 = 0$. Show that when gradient descent converges, it must converge to the least-norm solution $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x}\|_2^2$.

Proof:

(a) In class, we showed that if ∇F is Lipschitz with Lipschitz constant L, then choosing $\eta = 1/L$ guarantees the convergence of gradient descent. In particular, any $\eta \leq 1/L$ will guarantee the convergence of gradient descent, so we need only find L. We have

$$\begin{aligned} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|_2 &= \|2\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) - 2\mathbf{A}^\top(\mathbf{A}\mathbf{y} - \mathbf{b})\|_2 \\ &= \|2\mathbf{A}^\top\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2 \\ &\leq 2\|\mathbf{A}^\top\mathbf{A}\| \cdot \|\mathbf{x} - \mathbf{y}\|_2 \end{aligned}$$

where $\|\cdot\|$ denotes the operator norm. Hence choosing $\eta \leq (2\|\mathbf{A}^\top\mathbf{A}\|)^{-1}$ will guarantee the convergence of gradient descent for any initialization.

(b) Let us first prove the hint, namely, that if $\mathbf{x} \in \operatorname{img} \mathbf{A}^\top$ (i.e. if $\mathbf{x}$ is in the rowspan of $\mathbf{A}$) then so is $A\mathbf{x} - \eta\nabla F(\mathbf{x})$. Suppose then that $\mathbf{x} = \mathbf{A}^\top\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^n$. Then

$$\begin{aligned} \mathbf{y} = \mathbf{A} \cdot \mathbf{A}^\top\mathbf{u} - \nabla F(\mathbf{A}^\top\mathbf{u}) &= \mathbf{A} \cdot \mathbf{A}^\top\mathbf{u} - \nabla 2A^\top(\mathbf{A}\mathbf{A}^\top\mathbf{u} - \mathbf{u}) \\ &= \mathbf{A}^\top u - 2\eta\mathbf{A}^\top\mathbf{A}\mathbf{A}^\top\mathbf{u} - 2\mathbf{A}^\top\mathbf{b} \\ &= \mathbf{A}^\top\left(\mathbf{u} - 2\eta\mathbf{A}\mathbf{A}^\top\mathbf{u} - 2\mathbf{b}\right) \implies \mathbf{y} \in \operatorname{img}\mathbf{A}^\top. \end{aligned}$$

Because the update rule is continuous and the image of affine linear transformations is closed, we can further conclude that an initialization $\mathbf{x}_0$ is in the rowspan of $\mathbf{A}$ if and only if the point $\mathbf{x}^*$ it converges to is in the rowspan of $\mathbf{A}$, provided η is chosen small enough to guarantee convergence.

Now we prove that any two points initialized in the rowspan of $\mathbf{A}$ converge to the same point. Take $\mathbf{x}_0 = \mathbf{A}^\top\mathbf{u}_0$ to be an initialization for some $\mathbf{u}_0 \in \mathbb{R}^n$. By what we have previously shown, $\mathbf{x}^* = \mathbf{A}^\top\mathbf{u}^*$ for some $\mathbf{u}^* \in \mathbb{R}^n$, supposing we have chosen η to be small enough. Since $\mathbf{x}^*$ is a stable point of the update rule, we get that $\nabla F(\mathbf{x}^*) = 0$ and hence

$$\begin{aligned}\nabla F(\mathbf{A}^\top \mathbf{u}^*) = 2\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top \mathbf{u}^* - \mathbf{b}) = 0 \\ \implies \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top \mathbf{u}^* - \mathbf{b}) = 0 \\ \implies \mathbf{A}^\top \mathbf{u}^* - \mathbf{b} = 0\end{aligned}$$

since $\mathbf{A}$ is full rank with $n < d$ (so $\ker A^\top = 0$). This means $\mathbf{u}^* = (\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{b}$, noting that the inverse $(\mathbf{A}\mathbf{A}^\top)^{-1}$ exists again because $\mathbf{A}$ is fully rank with $n < d$. Using this expression for $\mathbf{u}^*$ gives us that $\mathbf{x}^* = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{b}$, which notably does not depend on the initialization, implying that any two points initialized in the rowspan of $\mathbf{A}$ converge to the same point.

Finally, consider two different initialization $\mathbf{y}_0 \in \mathbb{R}^d \setminus \operatorname{img}(A^\top)$ and $\mathbf{x}_0 \in \operatorname{img}(A^\top)$. As before, $\mathbf{x}_0 = \mathbf{A}^\top \mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^n$. Since $\mathbf{y}_0$ is not in the rowspan of $\mathbf{A}$, the stable point $\mathbf{y}^*$ of the update rule to which $\mathbf{y}_0$ converges is also not in $\operatorname{img}(\mathbf{A}^\top)$. Hence $\mathbf{y}^* = \mathbf{A}^\top \mathbf{u}^* + \mathbf{v}$ for some $\mathbf{v} \notin \operatorname{img} \mathbf{A}^\top$, where $\mathbf{u}^*$ is as above. The stability condition $\nabla F(\mathbf{y}^*) = 0$ gives us $\mathbf{A}\mathbf{A}^\top \mathbf{u}^* + \mathbf{A}\mathbf{v} - \mathbf{b} = 0$ repeating the calculation from the last paragraph. But $\mathbf{A}\mathbf{A}^\top \mathbf{u}^* - \mathbf{b} = 0$, so $\mathbf{A}\mathbf{v} = 0$. This means

$$\begin{aligned}\|\mathbf{y}^*\|_2^2 &= (\mathbf{A}^\top \mathbf{u}^* + \mathbf{v})^\top(\mathbf{A}^\top \mathbf{u}^* + \mathbf{v}) \\ &= \mathbf{u}^\top \mathbf{A}^\top \mathbf{A}\mathbf{u} + \mathbf{u}^\top \mathbf{A} v + (\mathbf{A}\mathbf{v})^\top \mathbf{u} + \mathbf{v}^\top \mathbf{v} \\ &= \mathbf{u}^\top \mathbf{A}^\top \mathbf{A}\mathbf{u} + 0 + 0 + \mathbf{v}^\top \mathbf{v} \\ &\geq \mathbf{u}^\top \mathbf{A}^\top \mathbf{A}\mathbf{u} \\ &= \|\mathbf{x}^*\|_2^2.\end{aligned}$$

Thus, $\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x}\|_2^2$. Any point in the rowspan of $\mathbf{A}$ converges to $\mathbf{x}^*$ under the update rule; in particular, the initialization $\mathbf{x}_0 = 0$ converges to $\mathbf{x}^*$, proving the desired result.

□

EXERCISE 2. Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a differentiable function. It satisfies the PL-inequality if there exists a constant $\mu > 0$ such that for all $w \in \mathbb{R}^d$ it holds

$$\frac{1}{2}\|\nabla f(w)\|_2^2 \geq \mu(f(w) - f^*).$$

By contrast we say f is *invex* if there exists a function $\eta : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ such that for all $x, y \in \mathbb{R}^d$ it holds

$$f(y) \geq f(x) + \nabla f(x)^\top \eta(x, y).$$

(a) Show that if f satisfies the PL-inequality then f is invex.

(b) Show that any stationary point of an invex function is a global minimizer.

Proof:

(a) Since f satisfies the PL-inequality, for some $\mu > 0$

$$\frac{1}{2}\|\nabla f(w)\|_2^2 \geq \mu(f(w) - f^*)$$

for all $w \in \mathbb{R}^d$, where f^* is the global minimum of f. Rearranging, we get

$$\begin{aligned} & \frac{1}{2}\|\nabla f(w)\|_2^2 \geq \mu(f(w) - f^*) \\ \implies & \nabla f(w)^\top \nabla f(w) \geq 2\mu f(w) - 2\mu f^* \\ \implies & -\nabla f(w)^\top \nabla f(w) \leq 2\mu f^* - 2\mu f(w) \leq f(w) \leq 2\mu f(u) - 2\mu f(w) \end{aligned}$$

for any $w, u \in \mathbb{R}^d$, since f^* is the global minimum of f. This in turn implies that

$$2\mu f(w) - \nabla f(w)^\top \nabla f(w) \leq 2\mu f(u),$$

so if we set $\eta(x, y) = -\frac{1}{2\mu}\nabla f(x)$ then we get

$$f(x) = \nabla f(x)^\top \cdot \eta(x, y) \leq f(y)$$

for all $x, y \in \mathbb{R}^d$. This proves that f is invex.

(b) A point $\mathbf{x}$ is a stationary point of f if $\nabla f(x) = 0$. If f is invex, then we get

$$f(y) \geq f(x) + \nabla f(x)^t op\eta(x, y) = f(x)$$

for all $y \in \mathbb{R}^d$. Hence any stationary point of f is a global minima. Combining with part (a) we see that any function which satisfies the PL-inequality is easily optimized.

□

EXERCISE 3. In your favorite programming language, implement stochastic gradient-descent for the linear least squares loss $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{AW} - \mathbf{b}\|_2^2$. Provide convergence plots to validate the convergence guarantees for SGD discussed in class. Specifically, compare empirical and theoretical convergence rates when $\mathbf{A} \in \mathbb{R}^{10{,}000\times 1{,}000}$ has iid $\mathcal{N}(0, 1/\sqrt{1000})$ Gaussian entries and $\mathbf{b} = \mathbf{A1} + \varepsilon$ where $\mathbf{1}$ is the all-ones vector and ε has iid Gaussian antries with variance 1, then 0.1, then 0.01 and finally 0. Repeat the comparisons but now consider $\mathbf{A} \in \mathbb{R}^{10000\times 1000}$ whose jth row has iid $\mathcal{N}(0, 1/\sqrt{1000j})$ Gaussian entries. (*Note your answer should include 8-plots, because there are two choices of* $\mathbf{A}$ *and four different choices of* ε.)

EXERCISE 4. Consider a three-state Markov chain with stationary probabilities $\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right)$. consider the Metropolis-Hastings algorithm with G the complete graph on these three vertices. For each edge and each direction, what is the expected probability that we would actually make a move along the edge?

Proof: Recall that the Metropolis transition probabilities are

$$p_{xy} = \frac{1}{r}\min\left(1, \frac{\pi(y)}{\pi(x)}\right)$$

if x and y are distinct but adjacent and

$$p_{xx} = 1 - \sum_{y \neq x} p_{xy}.$$

Let a, b and c be the vertices of the graph. Then

$$\begin{aligned} p_{ab} &= \frac{1}{2}\frac{2}{1} \cdot \frac{1}{3} = \frac{1}{3} \\ p_{ac} &= \frac{1}{2}\frac{2}{1} \cdot \frac{1}{6} = \frac{1}{6} \\ p_{aa} &= 1 - \frac{1}{3} - \frac{1}{6} = \frac{1}{2}. \end{aligned}$$

The other transition probabilities are

$$p_{ba} = \frac{1}{2}, \qquad p_{bc} = \frac{1}{4}, \qquad p_{bb} = \frac{1}{4}$$

and

$$p_{ca} = \frac{1}{2}, \qquad p_{cb} = \frac{1}{3}, \qquad p_{cc} = \frac{1}{6}.$$

□

Exercise 5. Consider the probability distribution $p(\mathbf{x})$ where $\mathbf{x} \in \{0,1\}^{100}$ such that $p(0) = \frac{1}{2}$ and $p(\mathbf{x}) = \frac{1/2}{2^{100}-1}$. How does Gibbs sampling behave here?

Proof: The Gibbs transition probabilities are given by

$$p_{xy} = \begin{cases} \frac{1}{d}\pi(y_i \mid x_1, ..., \hat{x}_i, ..., x_d) & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ only in } i \\ 0 & \text{otherwise} \end{cases}.$$

Let $\hat{e}_i$ denote the element of $\{0,1\}^{100}$ whose ith component is 1 and is 0 elsewhere. We have three cases to examine.

If we are currently at $\mathbf{0}$, then

- there is a $\frac{1}{100}\frac{\frac{1/2}{2^{100}-1}}{\frac{1}{2}+\frac{1/2}{2^{100}-1}} \approx \frac{1}{100} \cdot \frac{1}{2^{100}-1}$ chance of moving to $\hat{e}_i$ for any $i \in \{1, ..., 100\}$. Altogether, we have a 1 in $2^{100} - 1$ chance of leaving 0 at all.
- We have a $1 - \frac{1}{2^{100}-1} \approx 1$ chance of remaining at zero.

Hence, if we ever reach 0 then we will stay at zero, since $2^{100} - 1$ is a huge number.

If we are currently at $\hat{e}_i$, then we

- have a $\frac{1}{100}\frac{1/2}{\frac{1}{2}+\frac{1/2}{2^{100}-1}} \approx \frac{1}{100}$ chance of moving to 0.
- have a $\frac{1}{100}\frac{\frac{1/2}{2^{100}-1}}{\frac{1/2}{2^{100}-1}+\frac{1/2}{2^{100}-1}} \approx \frac{1}{200}$ chance of moving to some other nonzero point, of which there are 99 adjacent to $\hat{e}_i$ giving us an approximately $\frac{1}{2}$ chance point of moving to a point which is not 0

- have an approximately $1 - \frac{1}{100} - \frac{1}{2} = \frac{1}{2} - \frac{1}{100}$ chance of remaining at $\hat{e}_i$.

If we are currently at $\mathbf{x} \neq \hat{e}_i, \mathbf{0}$, then we

- have a $\frac{1}{100} \frac{\frac{1/2}{2^{100}-1}}{\frac{1/2}{2^{100}-1} + \frac{1/2}{2^{100}-1}} \approx \frac{1}{200}$ chance of moving to any individual neighbor of $\mathbf{x}$, or altogether a 1 in 2 chance of leaving $\mathbf{x}$ to *some* other point
- have an $1 - 100 \cdot \frac{1}{200} \approx \frac{1}{2}$ chance of remaining at $\mathbf{x}$.

If we initialize a random walk on G at $\mathbf{0}$ then we will remain there functionally forever. If we initialize it at any other point, then we have a $\frac{1}{2}$ chance to leave and a $\frac{1}{2}$ chance to remain. The situation is slightly different at a point neighboring $\mathbf{0}$, where we have twice the chance of transitioning to $\mathbf{0}$ than to any other point. Thus, a random walk on G will visit a variety of points, transitioning to a new point every 2 steps on average, unless it reaches $\mathbf{0}$, in which case it will remain there indefinitely. However, the chance of reaching $\mathbf{0}$ from a random initialization is just as small as the chance of leaving $\mathbf{0}$, since there are 2^{100} points in total. $\square$