CPSC 486/586: Probabilistic Machine Learning	January 20, 2023

Lecture 2

Lecturer: Jun-Kun Wang | *Scribe: Sebastian Varma*

Optimization

1. Applications in ML

Optimization is used extensively throughout ML algorithms. For example, one can represent the weights in a linear classifier or a neural network as $w \in \mathbb{R}^d$. For any sort of optimization, we need to specify an objective function. Suppose we have n samples of the form (x_i, y_i) where the x_i are the features and the y_i are the labels. A typical optimization problem looks like

$$w^* = \arg\min_{w \in S} \frac{1}{n} \sum_{i=1}^{n} \ell(w; x_i, y_i)$$

for some set S and loss function ℓ which provides the loss for a single example (x_i, y_i) given the weight vector w. One such specific ℓ would be squared loss, in which case we would have:

$$\ell(w; x, y) = \frac{1}{2}(y - w^T x)^2$$

2. Gradient Descent

Suppose we have some differentiable $f : \mathbb{R}^d \to \mathbb{R}$ which we would like to minimize. The gradient descent algorithm for finding a minimum is as follows. Initialize $w_1 \in \mathbb{R}^d$. For $k = 1, 2, ...$, for $\eta > 0$ which is step size / learning rate, set:

$$w_{k+1} = w_k - \eta \nabla f(w_k)$$

We denote here $\nabla f(w)$ as the gradient of f at w, which is defined as

$$\nabla f(w) = \begin{pmatrix} \frac{\partial f(w)}{\partial w_1} \\ \frac{\partial f(w)}{\partial w_2} \\ \vdots \\ \frac{\partial f(w)}{\partial w_d} \end{pmatrix} \in \mathbb{R}^d$$

We explain now why it is that we choose to move in the direction which is opposite the gradient. We want it to be the case that $\frac{\partial f(W_t)}{\partial t} \leq 0$ so that our algorithm finds smaller and smaller W_t as the algorithm progresses. By the chain rule, we calculate:

$$\frac{\partial f(W_t)}{\partial t} = \left\langle \nabla f(W_t), \frac{dW_t}{dt} \right\rangle = \sum_{i=1}^{d} \frac{\partial f(W)}{\partial W_t(i)} \cdot \frac{dW_t(i)}{dt}$$

We consider now two options for setting $\frac{dW_t}{dt}$ to provide some motivation for why we choose to set it to the negative gradient.

$$\text{Option 1}: \frac{dW_t}{dt} = \nabla f(W_t)$$

$$\text{Option 2}: \frac{dW_t}{dt} = -\nabla f(W_t)$$

Under option 1, we calculate:

$$\frac{\partial f(W_t)}{\partial t} = \langle \nabla f(W_t), \frac{dW_t}{dt} \rangle = \langle \nabla f(W_t), \nabla f(W_t) \rangle = \left\| \nabla f(W_t) \right\|^2 \geq 0$$

Under option 2, we calculate:

$$\frac{\partial f(W_t)}{\partial t} = \langle \nabla f(W_t), \frac{dW_t}{dt} \rangle = \langle \nabla f(W_t), -\nabla f(W_t) \rangle = -\left\| \nabla f(W_t) \right\|^2 \leq 0$$

Thus, under option 2, we have the obtained result where we decrease f in time. Option 2 is called gradient flow. It is a continuous-time algorithm. Gradient descent (GD) is a discrete-time algorithm because the indices on the W are countable rather than uncountable as in gradient flow.

Lemma 1. *GF is GD where $\eta \to 0$.*

Proof. From the definition of GD,

$$\frac{W_{k+1} - W_k}{\eta} = -\nabla f(W_k)$$

Taking the limit as $\eta \to 0$ and using the definition of a derivative:

$$\lim_{\eta \to 0} \frac{W_{k+1} - W_k}{\eta} = \lim_{\eta \to 0} -\nabla f(W_k)$$

$$\lim_{\eta \to 0} \frac{W_{k+1} - W_k}{\eta} = -\nabla f(W_k)$$

$$\frac{\partial W_t}{\partial t} = -\nabla f(W_t)$$

□

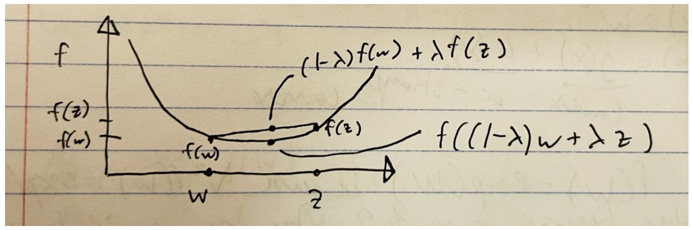
Classes of functions

1. Convex functions

Definition 1 (Convexity)**.** *f is convex if for any $w, z \in \mathbb{R}^d$ and $\lambda \in (0, 1)$,*

$$f((1-\lambda)w + \lambda z) \leq (1-\lambda)f(w) + \lambda f(z)$$

Remark: This definition says that if we draw a chord between two points, the function f will be below the chord on the open interval between the two points, as shown in the image below:



Two equivalent definitions are given below:

Definition 2 (Convexity). *f is convex if* $f(w) \geq f(z) + \langle \nabla f(z), w - z \rangle \; \forall w, z \in \mathbb{R}^d$.

Definition 3 (Convexity). *f is convex if, for twice-differentiable f,* $\lambda_{min}(\nabla^2 f(w)) \geq 0$.

Remark: Examples of convex functions include square loss functions such as $f(w) = \frac{1}{2}(y_i - W^T x_i)^2$ or hinge loss functions.

2. Strongly convex functions

Definition 4 (Strongly convex). *f is $\alpha-$strongly convex if $\forall w, z$ and $\alpha > 0$,*

$$f(w) \geq f(z) + \langle \nabla f(z), w - z \rangle + \frac{\alpha}{2}\|w - z\|^2$$

Remark: Since $\frac{\alpha}{2} > 0$ and norms are non-negative, the last term on the RHS of the defining criterion is non-negative, so then by definition 2, strongly convex implies convex.

We give below an equivalent definition of strongly convex

Definition 5 (Strongly convex). *f is $\alpha-$strongly convex if $\forall w, z$ and $\alpha > 0$,*

$$\langle \nabla f(w) - \nabla f(z), w - z \rangle \geq \alpha \|w - z\|^2$$

We prove one direction of this equivalence:

Lemma 2. *Definition 4 implies Definition 5*

Proof. Applying definition 4 twice (the latter time swapping w and z) gives:

$$f(w) \geq f(z) + \langle \nabla f(z), w - z \rangle + \frac{\alpha}{2}\|w - z\|^2$$

$$f(z) \geq f(w) + \langle \nabla f(w), z - w \rangle + \frac{\alpha}{2}\|w - z\|^2$$

Summing these inequalities:

$$f(z) + f(w) \geq f(z) + f(w) + \langle \nabla f(z), w - z \rangle + \langle \nabla f(w), z - w \rangle + 2 \cdot \frac{\alpha}{2}\|w - z\|^2$$

$$0 \geq \langle \nabla f(z), w - z \rangle + \langle \nabla f(w), z - w \rangle + \alpha \|w - z\|^2$$

$$\langle \nabla f(w) - \nabla f(z), w - z \rangle \geq \alpha \|w - z\|^2$$

$\square$

Finally, we give another equivalent definition:

Definition 6 (Strongly convex). *If f is twice differentiable, f is strongly convex when*

$$\lambda_{min}(\nabla^2 f(x)) \geq \alpha > 0$$

Examples: $f(w) = w^2$ is strongly convex. $F(x) = g(x) + h(x)$ is strongly convex for convex g and strongly convex h.

Counterexample: We claim $f(w) = \exp(w)$ is not strongly-convex because $\nabla^2 f(w) = \exp(w)$, and when $w \to -\infty$ the hessian goes to 0. Then the hessian cannot be lower-bounded by a strictly positive number.

3. Non-convex functions

Definition 7 (Non-convex functions). *Any f which is not convex belongs to this class.*

Remark: Minimizing non-convex functions is generally much harder than minimizing convex functions, as we cannot depend upon certain nice properties like with convex functions.

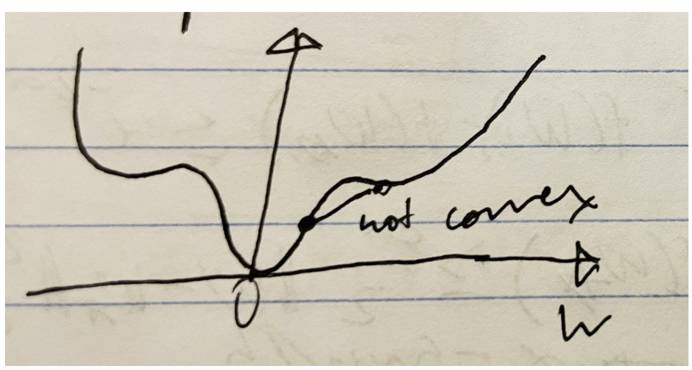
4. Polyak-Łojasiewicz Functions (PL condition)

Definition 8 ($\mu-$PL Functions). *A function f is $\mu-$PL globally if for any point $w \in \mathbb{R}^d$ it satisfies the PL condition (also called the gradient domination condition), which is as follows:*

$$\left\|\nabla f(w)\right\|^2 \geq 2\mu(f(w) - \min_{w'} f(w'))$$

for $\mu > 0$.

Example: The function $f(w) = w^2 + 4\sin^2(w)$ is shown below:



It can be shown that this function is a PL function. However, as shown by the chord which is drawn and falls below the function, this function is not convex. This example demonstrates that PL functions are not necessarily convex.

Other examples: The objective function of an ultra-wide network, the policy gradient in reinforcement learning, and $f(w) = \frac{1}{2}w^T Aw$ where A is positive semi-definite (so $\lambda_{min}(A) \geq 0$) are all PL functions. Note that the last example is not strongly convex because $\nabla^2 f = A$ and λ_{min} is not necessarily positive, as it could be zero.

Some results related to these classes of functions

Throughout, let $w^* \in \arg\min_w f(w)$.

Lemma 3. *$\alpha-$strongly convex implies $\alpha - PL$.*

Proof. Suppose f is $\alpha-$strongly convex. Using definition 4, we have:

$$f(w^*) \geq f(w)+\langle \nabla f(w), w^*-w\rangle+\frac{\alpha}{2}\|w-w^*\|^2 \implies f(w)-f(w^*) \leq \langle \nabla f(w), \underbrace{w-w^*}_{\text{sign change}} \rangle-\frac{\alpha}{2}\|w-w^*\|^2$$

$$\begin{aligned} f(w)-f(w^*) &\leq \langle \nabla f(w), w-w^*\rangle - \frac{\alpha}{2}\|w-w^*\|^2 \\ &= \langle \nabla f(w), w-w^*\rangle - \frac{\alpha}{2}\|w-w^*\|^2 - \frac{1}{2\alpha}\|\nabla f(w)\|^2 + \frac{1}{2\alpha}\|\nabla f(w)\|^2 \\ &= -\frac{1}{2}\left(\frac{1}{\alpha}\|\nabla f(w)\|^2 - 2\langle \nabla f(w), w-w^*\rangle + \alpha\|w-w^*\|^2\right) + \frac{1}{2\alpha}\|\nabla f(w)\|^2 \\ &= -\frac{1}{2}\left\|\sqrt{\alpha}(w-w^*) - \frac{1}{\sqrt{\alpha}}\nabla f(w)\right\|^2 + \frac{1}{2\alpha}\|\nabla f(w)\|^2 \end{aligned}$$

By the non-negativity of norms:

$$f(w)-f(w^*) \leq \frac{1}{2\alpha}\|\nabla f(w)\|^2 \implies \|\nabla f(w)\|^2 \geq 2\alpha(f(w) - \min_w f(w))$$

This is definition 8 for $\mu = \alpha$, so we are done. □

Theorem 1. *If f is $\mu - PL$, then the convergence of GF is upper-bounded as follows:*

$$f(W_t) - f(W^*) \leq e^{-2\mu t}(f(W_0) - f(W^*))$$

Proof. As stated earlier, the chain rule gives us (since $f(W^*)$ is a constant):

$$\frac{d(f(W_t) - f(W^*))}{dt} = \langle \nabla f(W_t), \frac{dW_t}{dt}\rangle$$

By the definition of GF:

$$\frac{d(f(W_t) - f(W^*))}{dt} = \langle \nabla f(W_t), -\nabla f(W_t)\rangle = -\|\nabla f(W_t)\|^2$$

Applying the PL condition:

$$\frac{d(f(W_t) - f(W^*))}{dt} \leq -2\mu(f(W_t) - f(W^*))$$

We now solve the general differential inequality $\frac{dA_t}{dt} \leq -2\mu A_t$ via separation of variables:

$$\frac{dA_t}{dt} \leq -2\mu A_t \implies \frac{dA_t}{A_t} \leq -2\mu dt \implies \int_0^t \frac{dA_t}{A_t} \leq \int_0^t -2\mu dt \implies \log A_t - \log A_0 \leq -2\mu t \implies A_t \leq A_0 \exp(-2\mu t)$$

We now plug in $A_t = f(W_t) - f(W^*)$:

$$f(W_t) - f(W^*) \leq e^{-2\mu t}(f(W_0) - f(W^*))$$

□

Definition 9 ($\alpha-$growth condition)**.** *The $\alpha-$growth condition for f is*

$$f(W) - f(W^*) \geq \frac{\alpha}{2}\left\|W - W^*\right\|^2$$

Lemma 4. *$\alpha-$strong convexity implies $\alpha-$growth.*

Proof. Suppose f is $\alpha-$strong convex. Using definition 4 and the fact that the gradient is zero at the minimum, we have

$$\begin{aligned} f(W) - f(W^*) &\geq \langle \nabla f(W^*), W - W^* \rangle + \frac{\alpha}{2}\left\|W - W^*\right\|^2 \\ &= \langle 0, W - W^* \rangle + \frac{\alpha}{2}\left\|W - W^*\right\|^2 \\ &= \frac{\alpha}{2}\left\|W - W^*\right\|^2 \end{aligned}$$

□

Lemma 5. *$\alpha-$strong convexity implies $\alpha-PL$, which implies $\alpha-$growth*

We have shown the first implication but not the second. For now it will be stated as a fact.

Theorem 2. *Assume f is $\alpha-$strongly convex. Then GF has exponential contraction. Specifically, if we have*

$$\frac{dW_t}{dt} = -\nabla f(W_t) \quad \textit{and} \quad \frac{dZ_t}{dt} = -\nabla f(Z_t)$$

then

$$\|W_t - Z_t\|^2 \leq e^{-2\alpha t}\|W_0 - Z_0\|^2$$

Proof. By the chain rule:

$$\frac{d}{dt}\|W_t - Z_t\|^2 = 2\langle W_t - Z_t, \frac{dW_t}{dt} - \frac{dZ_t}{dt} \rangle = 2\langle W_t - Z_t, -\nabla f(W_t) + \nabla f(Z_t) \rangle = -2\langle W_t - Z_t, \nabla f(W_t) - \nabla f(Z_t) \rangle$$

By definition 5 of strong convexity:

$$\frac{d}{dt}\|W_t - Z_t\|^2 \leq -2\alpha\|W_t - Z_t\|^2$$

This differential inequality is of the form $\frac{dA_t}{dt} \leq -2\alpha A_t$ which we have already solved, so letting $A_t = \|W_t - Z_t\|^2$, we obtain:

$$\|W_t - Z_t\|^2 \leq e^{-2\alpha t}\|W_0 - Z_0\|^2$$

□