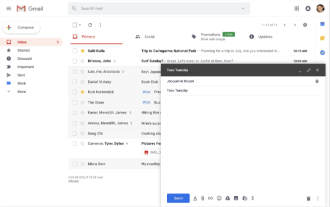
CMPT 409/981: Optimization for Machine Learning

Lecture 1

Sharan Vaswani

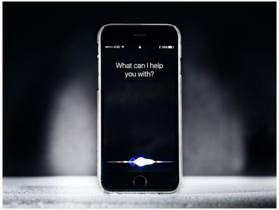
September 8, 2022

Successes of Machine Learning



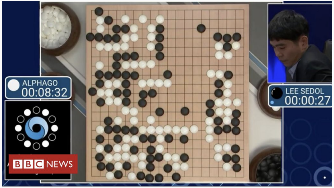
https://www.blog.google/products/gmail/subject-write-emails-faster-smart-compose-gmail/

(a) Natural language processing



https://www.cnet.com/news/what-is-siri/

(b) Speech recognition



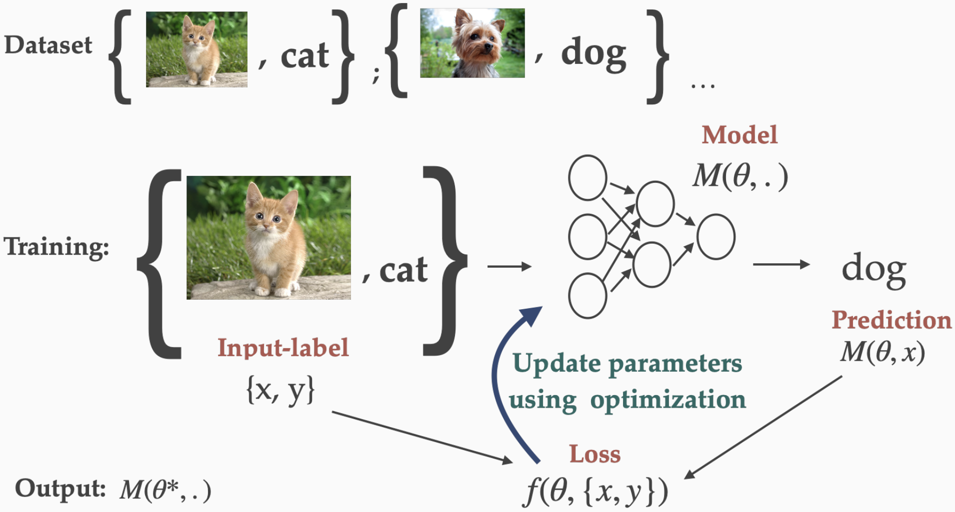
https://www.bbc.com/news/technology-35785875

(c) Reinforcement learning

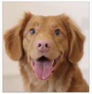
https://www.pbs.org/newshour/science/in-a-crash-should-self-driving-cars-save-passengers-or-pedestrians-2-million-people-weigh-in

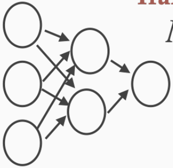
(d) Self-driving cars

Machine Learning 101



Machine Learning 101

Validation Dataset: $\{$  , cat $\}$; $\{$  , dog $\}$ …

Validation: $\{$  , dog $\}$ →  → dog

{x, y}

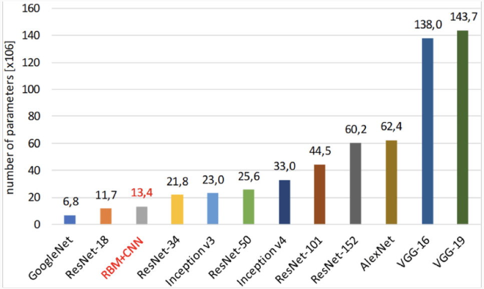
Trained Model
$M(\theta^*, .)$

Prediction
$M(\theta^*, x)$

Output: Validation Accuracy

← Measures how good the trained model is

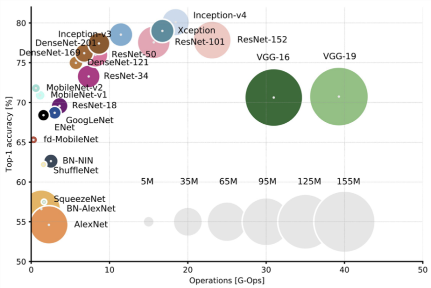
Modern Machine Learning



Sobczak, Szymon, et al. "Restricted Boltzmann machine as an aggregation technique for binary descriptors.", 2019.

Model size

(a)



Canziani et al, "An Analysis of Deep Neural Network Models for Practical Applications", 2016.

Number of operations for computing the loss

(b)

Figure 1: Models for multi-class classification on Image-Net. Number of examples = 1.2 M

Faster optimization methods can have a big practical impact!

Optimization for Machine Learning

- **(Non)-Convex minimization**: Supervised learning (classification/regression), Matrix factorization for recommender systems, Image denoising.
- **Online optimization**: Learning how to play Go/Atari games, Imitating an expert and learning from demonstrations, Regulating control systems like industrial plants.
- **Min-Max optimization**: Generative Adversarial Networks, Adversarial Learning, Multi-agent RL.

Course structure

Objective: Introduce foundational optimization concepts with applications to machine learning.

Syllabus:

- **(Non)-Convex minimization**: Gradient Descent, Momentum/Acceleration, Mirror Descent, Newton/Quasi-Newton methods, Stochastic gradient descent (SGD), Variance reduction
- **Online optimization**: Follow the (regularized) leader, Adaptive methods (AdaGrad, Adam)
- **Min-Max optimization**: (Stochastic) Gradient Descent-Ascent, (Stochastic) Extragradient

What we won't get time to cover in detail: Non-smooth optimization, Convex analysis, Global optimization.

What we won't get time to cover: Constrained optimization, Distributed optimization, Multi-objective optimization.

Course Logistics

- **Instructor**: Sharan Vaswani (TASC-1 8221) Email: `sharan_vaswani@sfu.ca`
- **Office Hours**: Monday 4 pm - 5 pm (TASC-1 8221), TBD
- **Teaching Assistant**: Zahra MiriKharaji Email: `zmirikha@sfu.ca`
- **Course Webpage**: `https://vaswanis.github.io/409_981-F22.html`
- **Piazza**: `https://piazza.com/sfu.ca/fall2022/cmpt409981/home`
- **Prerequisites**: Linear Algebra, Multivariable calculus, (Undergraduate) Machine Learning

Course Logistics – Grading

Assignments $[4 \times 12.5\% = 50\%]$

- Assignments to be submitted online, typed up in Latex with accompanying code submitted as a zip file.
- Each assignment will be due in 10 days (at 11.59 pm PST).
- For some flexibility, each student is allowed 1 late-submission and can submit in the next class (no late submissions beyond that).
- If you use up your late-submission and submit late again, you will lose 50% of the mark.

Course Logistics – Grading

Final Project [50%]

- Aim is to give you a taste of research in Optimization.
- Projects to be done in groups of 3-4 (more details will be on Piazza)
- Will maintain a list on Piazza on possible project topics. You are free to choose from the list or propose a topic that combines Optimization with your own research area.
- Project Proposal [10%] – Discussion (before 20 October) + Report (due 24 October)
- Project Milestone [5%] – Update (before 20 November)
- Project Presentation [10%] (6 December)
- Project Report [25%] (15 December)

Questions?

Minimizing functions

Consider minimizing a function over the domain $\mathcal{D}$

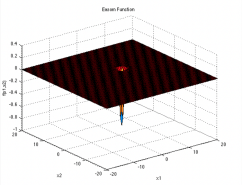
$$\min_{w \in \mathcal{D}} f(w).$$

Setting: Have access to a **zero-order oracle** – querying the oracle at $w \in \mathcal{D}$ returns $f(w)$.

Objective: For a target accuracy of $\epsilon > 0$, if $w^* \in \mathcal{D}$ is the minimizer of f, return a point $\hat{w} \in \mathcal{D}$ s.t. $f(\hat{w}) - f(w^*) \leq \epsilon$. Characterize the required number of oracle calls.

Example 1: Minimize a one-dimensional function s.t. $f(w) = 0$ for all $x \neq w^*$, and $f(w^*) = -\epsilon$.

Example 2: Easom function:
$f(x_1, x_2) = -\cos(x_1)\ -\cos(x_2)\ \exp(-(x_1 - \pi)^2 - (x_2 - \pi)^2)$.



Minimizing generic functions is hard! We need to make assumptions on the structure.

Lipschitz continuous functions

Consider minimizing a function over the domain $\mathcal{D}$:

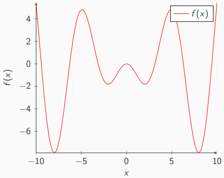
$$\min_{w \in \mathcal{D}} f(w).$$

Assumption: f is *Lipschitz continuous* meaning that f can not change arbitrarily fast as w changes. Formally, for any $x, y \in \mathcal{D}$,

$$|f(x) - f(y)| \leq G \, \|x - y\|$$

where G is the Lipschitz constant.

Example: $f(x) := -x \sin(x)$ in the $[-10, 10]$ interval.



Lipschitz continuity of the function immediately implies that the gradients are *bounded* i.e. for all $x \in \mathcal{D}$, $\|\nabla f(x)\| \leq G$.

Global Minimization

Consider minimizing a G-Lipschitz continuous function over a unit hyper-cube:

$$\min_{w \in [0,1]^d} f(w).$$

Objective: For a target accuracy of $\epsilon > 0$, if $w^* \in [0,1]^d$ is the minimizer of f, return a point $\hat{w} \in [0,1]^d$ s.t. $f(\hat{w}) - f(w^*) \leq \epsilon$. Characterize the required number of zero-order oracle calls.

Naive algorithm: Divide the hyper-cube into cubes with length of each side equal to $\epsilon' > 0$ (to be determined). Call the zero-order oracle on the centers of these $\frac{1}{(\epsilon')^d}$ cubes and return the point $\hat{w}$ with the minimum function value.

Analysis: The minimizer lies in/at the boundary of one of these cubes, and hence by returning the minimum $\hat{w}$, we guarantee that $\hat{w}$ is at most $\sqrt{d}\epsilon'$ away from w^* i.e. $\|\hat{w} - w^*\| \leq \sqrt{d}\epsilon'$. By G-Lipschitz continuity, $f(\hat{w}) - f(w^*) \leq G \|\hat{w} - w^*\| \leq G\sqrt{d}\epsilon'$. For a target accuracy of ϵ, we can set $\epsilon' = \frac{\epsilon}{G\sqrt{d}}$. Hence, for this naive algorithm, total number of oracle calls $= \left(\frac{G\sqrt{d}}{\epsilon}\right)^d$.

Global Minimization

Consider minimizing a differentiable, G-Lipschitz continuous function over a unit hyper-cube:

$$\min_{w \in [0,1]^d} f(w).$$

Q: Suppose we do a random search over the cubes? What is the expected number of function evaluations?

Ans: The probability of finding the correct cube is $p := \epsilon'^d$. If X is a r.v. equal to 1 if we find the correct cube, then X follows a Geometric distribution. Hence, expected number of evaluations is $\frac{1}{p} = (\epsilon')^d = \left(\frac{\epsilon}{G\sqrt{d}}\right)^d$.

Is our naive algorithm good? Can we do better?

Lower-Bound: For minimizing a G-Lipschitz continuous function over a unit hyper-cube, any algorithm requires $\Omega\left(\left(\frac{G}{\epsilon}\right)^d\right)$ calls to the zero-order oracle.

Our naive-algorithm is *sub-optimal* by a factor of $O\left((\sqrt{d})^d\right)$.

Questions?

Smooth functions

Recall that Lipschitz continuous functions have bounded gradients i.e. $\|\nabla f(w)\| \leq G$ and can still include *non-smooth* (not differentiable everywhere) functions.

For example, $f(x) = |x|$ is 1-Lipschitz continuous but not differentiable at $x = 0$ and the gradient changes from -1 at 0^- to $+1$ at 0^+.

An alternative assumption that we can make is that f is *smooth* – it is differentiable everywhere and its gradient is Lipschitz-continuous i.e. it can not change arbitrarily fast.

Formally, the gradient ∇f is L-Lipschitz continuous if for all $x, y \in \mathcal{D}$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$$

where L is the Lipschitz constant of the gradient (also called the smoothness constant of f).

Q: Does Lipschitz-continuity of the gradient imply Lipschitz-continuity of the function? Ans: No, $\frac{x^2}{2}$ is 1-smooth but its gradient equal to x is unbounded over $\mathbb{R}$.

Smooth functions – Examples

If f is twice-differentiable and smooth, then for all $x \in \mathcal{D}$, $\nabla^2 f(x) \preceq L\, I_d$ i.e. $\sigma_{\max}[\nabla^2 f(x)] \leq L$ where $\sigma_{\max}$ is the maximum singular value.

Q: Does $f(x) = x^3$ have a Lipschitz-continuous gradient over $\mathbb{R}$? Ans: No, $f''(x) = 12x$ which is not bounded as $x \to \infty$

Q: Does $f(x) = x^3$ have a Lipschitz-continuous gradient over $[0, 1]$?

Ans: Yes, because $f''(x) = 12x$ is bounded on $[0, 1]$.

Q: The *negative entropy function* is given by $f(x) = x \log(x)$. Does it have a Lipschitz-continuous gradient over $[0, 1]$? Ans: No, $f''(x) = 1/x \to \infty$ as $x \to 0$.

Smooth functions – Examples

Linear Regression on n points with d features. Feature matrix: $X \in \mathbb{R}^{n \times d}$, vector of measurements: $y \in \mathbb{R}^n$ and parameters $w \in \mathbb{R}^d$.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{2} \|Xw - y\|^2$$

$$f(w) = \frac{1}{2} [w^\intercal (X^\intercal X) w - 2 w^\intercal X^\intercal y + y^\intercal y] \, ; \nabla f(w) = X^\intercal X w - X^\intercal y \, ; \nabla f(w) = X^\intercal X$$

If f is L-smooth, then, $\sigma_{\max}[\nabla^2 f(w)] \leq L$ for all w. Hence, for linear regression $L = \lambda_{\max}[X^\intercal X]$.

Q: Is the linear regression loss-function Lipschitz continuous? Ans: No. Since $\|\nabla f(w)\| \to \infty$ as $w \to \infty$.

Q: Compute L for *ridge regression* – ℓ_2-regularized linear regression where $f(w) := \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$. Ans: $L = \lambda_{\max}[X^\intercal X] + \lambda$

Smooth functions

Claim: For an L-smooth function, $f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ for all $x, y \in \mathcal{D}$.
Proof:

$$f(y) = f(x) + \int_{t=0}^{1} [\nabla f(x + t(y - x))] (y - x)^\intercal \, dt \qquad \text{(Fundamental theorem of calculus)}$$

$$= f(x) + \langle \nabla f(x), y - x \rangle + \int_{t=0}^{1} [\nabla f(x + t(y - x))] (y - x)^\intercal \, dt - [\nabla f(x)] (y - x)^\intercal$$

$$= f(x) + \langle \nabla f(x), y - x \rangle + \int_{t=0}^{1} [\nabla f(x + t(y - x)) - \nabla f(x)] (y - x)^\intercal \, dt$$

$$\le f(x) + \langle \nabla f(x), y - x \rangle + \int_{t=0}^{1} \|\nabla f(x + t(y - x)) - \nabla f(x)\| \, \|y - x\| \, dt$$

(Cauchy–Schwarz)

$$\le f(x) + \langle \nabla f(x), y - x \rangle + L \int_{t=0}^{1} \|x + t(y - x) - x\| \, \|y - x\| \, dt \qquad \text{(Lipschitz continuity)}$$

$$= f(x) + \langle \nabla f(x), y - x \rangle + L \, \|y - x\|^2 \int_{t=0}^{1} t \, dt = f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

Smooth functions

The inequality $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ can be interpreted as a *global* quadratic upper-bound on f at point x i.e. the bound holds for all $y \in \mathcal{D}$.

There are other related (not necessarily equivalent) ways to state the L-smoothness of f (you will need to prove these in Assignment 1).

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

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \|x - y\|^2$$

Questions?

Local Minimization

Even though f is smooth, it still includes functions with multiple local/global minimum and stationary points. Eg: $f(x) = -x \sin(x)$.

Consider minimizing a smooth function over $\mathbb{R}^d$ (unconstrained minimization)

$$\min_{w \in \mathbb{R}^d} f(w).$$

Since we have seen that global minimization can be impossible (without Lipschitz assumption on f) or the number of oracle calls can be exponential in d, let us aim to solve an easier problem.

Access to a **first-order oracle** – query the oracle at point w and it returns $f(w)$ and $\nabla f(w)$.

Objective: For a target accuracy of $\epsilon > 0$, return a point $\hat{w}$ s.t. $\|\nabla f(\hat{w})\|^2 \leq \epsilon$? Characterize the required number of oracle calls.

We only care about making the gradient small and finding an approximate stationary point.

Local Minimization

Recall that L-smoothness of f implies that $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$.

Idea: Since the RHS is a global upper-bound on the true function, instead of minimizing the function directly, let us minimize the upper-bound at x w.r.t y.

Setting the gradient of the RHS w.r.t y to zero, we obtain $\hat{y}$ as:

$$\nabla f(x) + L [\hat{y} - x] = 0 \implies \hat{y} = x - \frac{1}{L} \nabla f(x)$$

This is exactly the gradient descent update at x!

We can do this iteratively i.e. starting at w_0, form the upper-bound at w_0, minimize it by setting $w_1 = w_0 - \frac{1}{L} \nabla f(w_0)$, then form the quadratic upper-bound at w_1 and repeat. Continue to do this until we find a point $\hat{w}$ s.t. $\|\nabla f(\hat{w})\|^2 \leq \epsilon$ and terminate.

This is exactly the gradient descent procedure – move in the direction of the negative gradient ("downhill") with *step-size* η equal to to $1/L$. Formally, at iteration k, the GD update is:

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