10. Optimization: Convex optimization

Moving on

- **Part 1.** Approximation
  - For any function $g(\cdot)$, we can make a NN $f(\cdot)$ such that $\|f - g\| \leq \epsilon$
  - **Key factors.** Model size, smoothness of $g(\cdot)$, smoothness of activation $\sigma(\cdot)$

- **Part 2.** Optimization
  - By training with GD, the risk converges $\hat{R}(\hat{w}^{(t)}) - \hat{R}(\bar{w}) \leq \phi(t)$
  - By training with GD, the parameter converges $\|\hat{w}^{(t)} - \bar{w}\| \leq \psi(t)$
  - **Key factors.** Smoothness and convexity of $R(\cdot)$, step size, …

Optimization

- In ML, we are trying to find nice ways to solve and analyze

$$\min_{w} \hat{R}(w)$$

  - Typically, $\hat{R}(w)$ is the training risk

$$\hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

  - However, will usually encapsulate everything as $\hat{R}(\cdot)$

(we'll use w instead of $\mathbf{w}$ to lower the chance of typo)

Optimization

- **Focus.** We analyze how the first-order algorithms work
  - Gradient descent

$$w_{t+1} = w_t - \eta \cdot \nabla \hat{R}(w)$$

  - Gradient flow

$$\dot{w}(t) = -\nabla \hat{R}(w)$$

- **This week.** Heavy assumptions, no neural nets

Smoothness

Smoothness

- Two assumptions make it easy: Smoothness & Convexity

Definition (Smoothness).

A function $\hat{R}$ is β-smooth whenever

$$\|\nabla \hat{R}(w) - \nabla \hat{R}(v)\| \leq \beta \|w - v\|, \qquad \forall w, v$$

- Exercise. How smooth is the case of linear regression, i.e.,

$$\hat{R}(w) = \|y - w^\top X\|^2$$

- ReLU networks are not smooth in general — but we can still draw some insights from smooth cases

Convex upper bound

- Given the smoothness, we can prove that there exists a convex upper bound on risk

Lemma (Convex upper bound).
Suppose that $\hat{R}$ is β-smooth. Then, we have

$$\hat{R}(v) \leq \hat{R}(w) + \langle \nabla \hat{R}(w), v - w \rangle + \frac{\beta}{2}\|w - v\|^2, \qquad \forall w, v$$

- **Meaning.** A smooth function cannot grow faster than quadratic functions ✏️

- **Remark.** Gradient descent can be viewed as minimizing this convex UB
  - when $1/\beta$ is the step size

Convex upper bound

$$\hat{R}(v) \leq \hat{R}(w) + \langle \nabla \hat{R}(w), v - w \rangle + \frac{\beta}{2}\|w - v\|^2, \qquad \forall w, v$$

- **Proof idea.** Consider a curve $t \mapsto \hat{R}(w + t(v - w))$ ✏️

GD reduces the risk

- Using the convex UB, we can show that GD always reduces the risk

Lemma (GD reduces the risk)

Let w_1 be a one-step-GD-updated version of w_0, i.e.,

$$w_1 = w_0 - \eta \cdot \nabla \hat{R}(w_0)$$

Then, we have

$$\hat{R}(w_1) \leq \hat{R}(w_0) - \eta \left(1 - \frac{\beta\eta}{2}\right) \|\nabla \hat{R}(w_0)\|^2$$

- This holds for any η
  - Select "useful" values of η

GD reduces the risk

$$w_1 = w_0 - \eta \cdot \nabla \hat{R}(w_0)$$

$$\hat{R}(w_1) \leq \hat{R}(w_0) - \eta \left(1 - \frac{\beta\eta}{2}\right) \|\nabla \hat{R}(w_0)\|^2$$

- **Proof idea.** Use the convex UB

$$\hat{R}(v) \leq \hat{R}(w) + \langle \nabla \hat{R}(w), v - w \rangle + \frac{\beta}{2}\|w - v\|^2, \qquad \forall w, v$$

GD reduces the risk

$$\hat{R}(w_1) \leq \hat{R}(w_0) - \eta \left(1 - \frac{\beta\eta}{2}\right) \|\nabla \hat{R}(w_0)\|^2$$

- **Remark.** Sometimes, useful to turn this into a form

$$\|\nabla \hat{R}(w_0)\|^2 \leq \frac{2}{\eta(2 - \beta\eta)} \left(\hat{R}(w_0) - \hat{R}(w_1)\right)$$

  - LHS:
    - Scale of the gradient
    - Scale of the parameter update
  - RHS:
    - Scale of the current risk
    - Scale of the risk decrement

GD does not go far

- Extending this idea, we can prove that the GD arrives at some stationary-ish point

Theorem 7.1.

Let w_t be a t-step updated version of w_0, with the learning rate $\eta \leq 2/\beta$. Then, we have

$$\min_{i<t} \|\nabla \hat{R}(w_i)\|^2 \leq \frac{2}{t\eta(2-\eta\beta)} \left(\hat{R}(w_0) - \inf_w \hat{R}(w)\right)$$

- **Remark.** Plugging in $\eta = 1/\beta$, we get simply

$$\min_{i<t} \|\nabla \hat{R}(w_i)\|^2 \leq \frac{2\beta}{t} \left(\hat{R}(w_0) - \inf_w \hat{R}(w)\right)$$

GD does not go far

$$\min_{i<t} \|\nabla \hat{R}(w_i)\|^2 \leq \frac{2}{t\eta(2-\eta\beta)} \left(\hat{R}(w_0) - \inf_w \hat{R}(w) \right)$$

- **Proof idea.**
  - Note that min $\leq$ avg
  - Invoke the previous property

$$\|\nabla \hat{R}(w_0)\|^2 \leq \frac{2}{\eta(2-\beta\eta)} \left(\hat{R}(w_0) - \hat{R}(w_1) \right)$$

GD does not go far

- We can prove similar results for the gradient flow

$$\inf_{s \in [0,t]} \|\nabla \hat{R}(w(s))\|^2 \leq \frac{1}{t}\left(\hat{R}(w(0)) - \hat{R}(w(t))\right)$$

- **Remark.** Much cleaner form
    - No η
    - No β

GD does not go far

$$\inf_{s\in[0,t]} \|\nabla \hat{R}(w(s))\|^2 \leq \frac{1}{t}\left(\hat{R}(w(0)) - \hat{R}(w(t))\right)$$

- **Proof idea.** Use the fact that

$$\hat{R}(w(t)) - \hat{R}(w(0)) = \int_0^t \langle \nabla \hat{R}(w(s)), \dot{w}(s)\rangle \, \mathrm{d}s$$

Convexity

Convexity

- Now, let's think about another good tool — convexity

Definition (Convex).

A differentiable function $\hat{R}$ is convex, whenever

$$\hat{R}(w') \geq \hat{R}(w) + \langle \nabla \hat{R}(w), w' - w \rangle$$

- **Remarks.**
  - Not the general definition, but useful one under differentiability
  - Synergy with smoothness
    - Recall the convex UB, which is a consequence of smoothness

$$\hat{R}(w') \leq \hat{R}(w) + \langle \nabla \hat{R}(w), w' - w \rangle + \frac{\beta}{2}\|w - w'\|^2, \qquad \forall w, v$$

Risk convergence

- Let's bring that synergy into action

Theorem 7.3.

Suppose that $\hat{R}$ is convex and β-smooth. Then, by GD with $\eta = 1/\beta$ we get: For any z, we have

$$\hat{R}(w_t) - \hat{R}(z) \leq \frac{\beta}{2t}\left(\|w_0 - z\|^2 - \|w_t - z\|^2\right).$$

- We can select the "reference point" z freely
    - what is the most useful choice? 🙋

Risk convergence

- Let's bring that synergy into action

Theorem 7.3.

Suppose that $\hat{R}$ is convex and β-smooth. Then, by GD with $\eta = 1/\beta$ we get: For any z, we have

$$\hat{R}(w_t) - \hat{R}(z) \leq \frac{\beta}{2t}\left(\|w_0 - z\|^2 - \|w_t - z\|^2\right).$$

- We can select the "reference point" z freely
  - what is the most useful choice? 🙋
  - **Answer.** Of course, select $z = \arg\min_z \hat{R}(z)$
    - Otherwise, LHS can be meaningless
  - This leads to the GD risk convergence at rate $\sim 1/t$

Risk convergence

$$\hat{R}(w_t) - \hat{R}(z) \leq \frac{\beta}{2t}\left(\|w_0 - z\|^2 - \|w_t - z\|^2\right).$$

- **Proof idea.** Use the decomposition

$$\|w_{i+1} - z\|^2 = \|w_i - z\|^2 - \frac{2}{\beta}\langle \nabla \hat{R}(w_i), w_i - z\rangle + \frac{1}{\beta^2}\|\nabla \hat{R}(w_i)\|^2$$

Risk convergence

- Rephrasing, we get

$$\frac{2}{\beta}\langle \nabla \hat{R}(w_i), w_i - z\rangle - \frac{1}{\beta^2}\|\nabla \hat{R}(w_i)\|^2 = \|w_i - z\|^2 - \|w_{i+1} - z\|^2$$

  - Blue. Convexity implies

$$\hat{R}(w_i) \geq \hat{R}(z) + \langle \nabla \hat{R}(w), w_i - z\rangle$$

  - Red. Smoothness implies

$$\|\nabla \hat{R}(w_i)\|^2 \leq 2\beta\left(\hat{R}(w_i) - \hat{R}(w_{i+1})\right)$$

Risk convergence

- There is a similar version for GF

Theorem 7.4.

For any $z \in \mathbb{R}^d$, GF for a convex $\hat{R}$ satisfies

$$\hat{R}(w(t)) \le \hat{R}(z) + \frac{1}{2t}\left(\|w(0) - z\|^2 - \|w(t) - z\|^2\right)$$

- **Remark.**
  - Again, no β
  - Holds for general reference point

Risk convergence

$$\hat{R}(w(t)) \leq \hat{R}(z) + \frac{1}{2t}\left(\|w(0)-z\|^2 - \|w(t)-z\|^2\right)$$

- **Proof.**

$$\begin{aligned}
\frac{1}{2}\left(\|w(t)-z\|^2 - \|w(0)-z\|^2\right) &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\|w(s)-z\|^2 \,\mathrm{d}s \\
&= \int_0^t \left\langle \frac{\mathrm{d}w}{\mathrm{d}s}, w(s)-z \right\rangle \mathrm{d}s \\
&= \int_0^t \left\langle \nabla \hat{R}(w(s)), z - w(s) \right\rangle \mathrm{d}s \\
&\leq \int_0^t \left(\hat{R}(z) - \hat{R}(w(s)) \right) \mathrm{d}s
\end{aligned}$$

Strong Convexity

Strong Convexity

- Consider a stronger assumption

Definition (Strongly convex).

A function $\hat{R}$ is λ-strongly convex whenever

$$\hat{R}(w') \geq \hat{R}(w) + \langle \nabla \hat{R}(w), w' - w \rangle + \frac{\lambda}{2}\|w' - w\|^2$$

- **Remark.** Even stronger synergy with the smoothness

$$\hat{R}(w') \leq \hat{R}(w) + \langle \nabla \hat{R}(w), w' - w \rangle + \frac{\beta}{2}\|w - w'\|^2, \qquad \forall w, v$$

Strong Convexity

- In fact, this is one of the reasons why we regularize

Proposition.

Suppose that $\hat{R}$ is convex. Then, the regularized risk $\hat{R}_{\text{reg}} := \hat{R} + \lambda \|w\|^2$ is 2λ-strongly convex

- **Proof idea.** Invoke the definition

$$\hat{R}(w') \geq \hat{R}(w) + \langle \nabla \hat{R}(w), w' - w \rangle + \frac{\lambda}{2} \|w' - w\|^2$$

Lower bound on the Gradient

- Strong convexity gives you a lower bound on the scale of the gradient

Lemma 7.1.

Suppose that $\hat{R}$ is λ-strongly convex. Then, for all w, we have

$$\hat{R}(w) - \inf_{v} \hat{R}(v) \leq \frac{1}{2\lambda} \|\nabla \hat{R}(w)\|^2$$

- **Remark.** Compare with the consequence of smoothness

$$\|\nabla \hat{R}(w)\|^2 \leq \frac{2}{\eta(2 - \beta\eta)} \left(\hat{R}(w) - \hat{R}(w')\right)$$

Lower bound on the Gradient

$$\hat{R}(w) - \inf_v \hat{R}(v) \leq \frac{1}{2\lambda} \|\nabla \hat{R}(w)\|^2$$

- **Proof idea.** For a fixed w, define a convex quadratic LB

$$Q_w(v) := \hat{R}(w) + \langle \nabla \hat{R}(w), v - w \rangle + \frac{\lambda}{2} \|v - w\|^2$$

  - Find the minimum $\hat{v}$
  - Then, do:

$$\inf_v \hat{R}(v) \geq \inf_v Q_w(v) = Q_w(\hat{v})$$

Risk convergence

- Given the strong convexity & smoothness, we can prove a stronger risk convergence bound

Theorem 7.5.

Suppose that $\hat{R}$ is λ-strongly convex and β-smooth, and let $\eta = 1/\beta$.
Then, for the risk minimizer $\bar{w}$, we have:

$$\hat{R}(w_t) - \hat{R}(\bar{w}) \leq (\hat{R}(w_0) - \hat{R}(\bar{w})) \cdot \exp\left(-\frac{t\lambda}{\beta}\right)$$

- **Note.** This is an exponential convergence
    - Much faster than $1/t$ for the convex case

Risk convergence

$$\hat{R}(w_t) - \hat{R}(\bar{w}) \leq (\hat{R}(w_0) - \hat{R}(\bar{w})) \cdot \exp\left(-\frac{t\lambda}{\beta}\right)$$

- **Proof sketch.**
  - Smoothness implies "GD reduces risk"

$$\hat{R}(w_{i+1}) - \hat{R}(\bar{w}) \leq \hat{R}(w_i) - \hat{R}(\bar{w}) - \frac{\|\nabla \hat{R}(w_i)\|^2}{2\beta}$$

  - Lemma 7.1. states

$$\hat{R}(w) - \inf_v \hat{R}(v) \leq \frac{1}{2\lambda}\|\nabla \hat{R}(w)\|^2$$

Parameter convergence

Theorem 7.5 (cont'd).

... and also, we have

$$\|w_t - \bar{w}\|^2 \leq \|w_0 - \bar{w}\|^2 \exp\left(\frac{-t\lambda}{\beta}\right)$$

- **Note.** The first parameter convergence guarantee

Parameter convergence

$$\|w_t - \bar{w}\|^2 \leq \|w_0 - \bar{w}\|^2 \exp\left(\frac{-t\lambda}{\beta}\right)$$

- **Proof idea.**
  - Let w' be an updated version of w
  - Then, we get

$$\|w' - \bar{w}\|^2 = \|w - \bar{w}\|^2 + \frac{2}{\beta}\langle \nabla \hat{R}(w), \bar{w} - w\rangle + \frac{1}{\beta^2}\|\nabla \hat{R}(w)\|^2$$

  - UB the 2nd term with the strong convexity
  - UB the 3rd term with the smoothness

Next up

- Polyak-Łojasiewicz condition
- Stochastic gradients