11. Optimization: Convex(?) optimization

Polyak-Łojasiewicz

Why is convexity useful?

- So far, we have seen that convexity + smoothness makes things easy

- If we have strong convexity, we have an LB on gradient

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

  - **Interpretation.** When suboptimal, GD updates rapidly

- This is paired with an UB on gradient for smooth functions

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

  - **Interpretation.** When near-optimal, GD updates small (as GD always reduces risk)

Polyak-Łojasiewicz

- In fact, it turns out that this gradient-risk bound is all we need

Definition (P-L condition).

A function $f(\cdot)$ is μ-PL whenever it satisfies:

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left(f(x) - \inf_x f(x)\right), \qquad \forall x$$

- We automatically have that a λ-strongly convex function is also λ-PL

- **Strong convexity.** Requires quadratic growth for any two points
- **PL condition.** Requires quadratic growth only around the optimum point

- Typically, our assumptions need to hold only locally (e.g., a ball containing initial point)

Polyak-Łojasiewicz

- In fact, it turns out that this gradient-risk bound is all we need

Proposition.

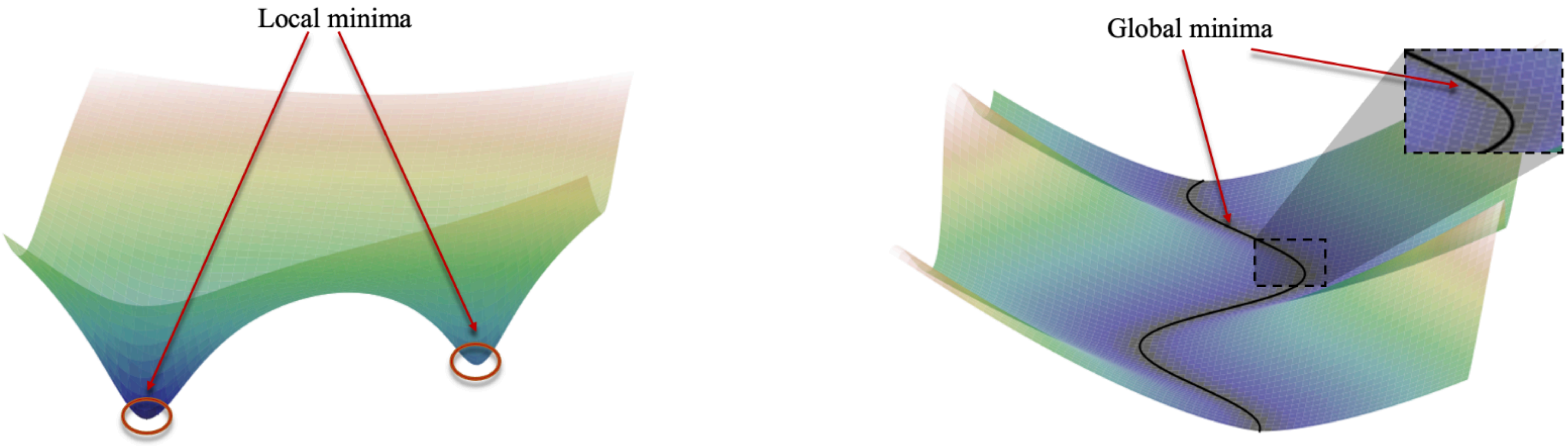
Suppose that $\hat{R}(\,\cdot\,)$ is μ-PL and β-smooth. Then, we have

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

- **Proof idea.** Same as in the strongly convex case!

Are neural net loss landscape PL?

- When sufficiently overparametrized, people argue that this is the case — c.f.,
    - Liu et al., “Loss landscapes and optimization in over-parameterized non-linear systems and neural networks,” Applied & Computational Harmonic Analysis, 2022
    - Islamov et al., “Loss Landscape Characterization of Neural Networks without Over-Parametrization,” NeurIPS 2024



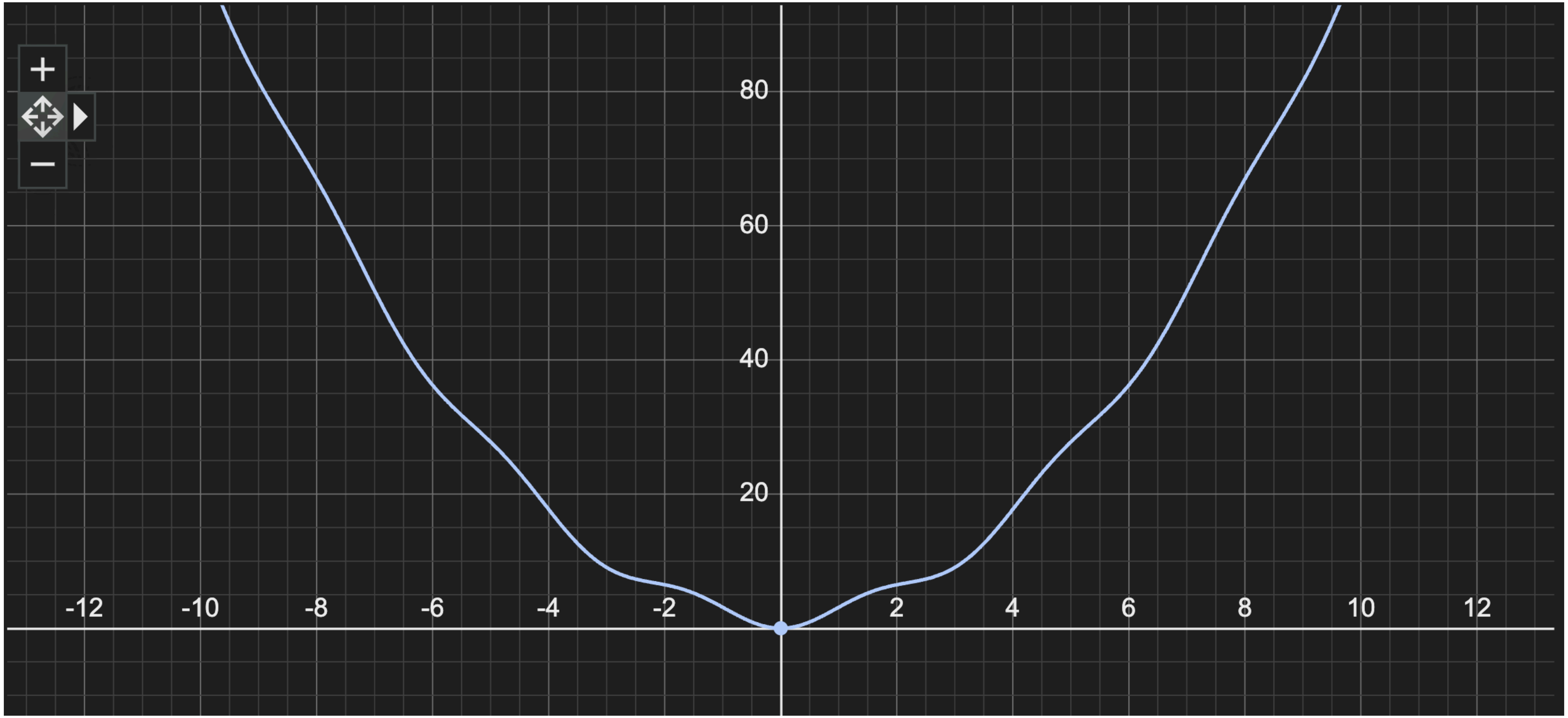
(a) Loss landscape of under-parameterized models

(b) Loss landscape of over-parameterized models

Figure 1: Panel (a): Loss landscape is locally convex at local minima. Panel (b): Loss landscape incompatible with local convexity as the set of global minima is not locally linear.

Examples

- Here are some examples of PL but nonconvex functions
- **Example.** $f(x) = x^2 + 3 \cdot \sin^2(x)$



Remarks

- There are many extensions and generalizations, for nonsmooth cases
    - Kurdyka-Łojasiewicz condition
    - α-β condition

Stochastic Gradients

Motivations

- We rarely use GD per se — instead, we use:
  - **SGD** (or mini-batch GD)
    - Memory. Need to store activations
    - Generalization. Large batch leads to suboptimal generalization
  - **Compressed Gradient**
    - Federated learning. Prune/Quantize
  - **Zeroth order optimization**
    - Black-box models. Proprietary models as a part of the pipeline
    - Computation. Does not require backward

Stochastic Gradients

- Formally, consider a generalized version of the gradient descent

$$w_{i+1} = w_i - \eta g_i$$

- Here, g_i is some estimate of the gradient $\nabla \hat{R}(w_i)$
    - Stochastic
    - Quantization noise
    - Sometimes, satisfies unbiasedness:

$$\mathbb{E}[g_i] = \nabla \hat{R}(w_i)$$

- **Goal.** Extend the usual analysis to analyze SGD
    - Risk convergence

Risk convergence

Lemma 7.2.

Suppose that $\hat{R}$ is convex, and let $G := \max_i \|g_i\|$. Let $\eta = 1/\sqrt{t}$. Then, for any z, we have

$$\frac{1}{t}\sum_{i<t}\hat{R}(w_i) \leq \hat{R}(z) + \frac{\|w_0 - z\|^2}{2\sqrt{t}} + \frac{G^2}{2\sqrt{t}} + \frac{1}{t}\sum_{i<t}\epsilon_i$$

where we use the shorthand $\epsilon_i = \langle g_i - \nabla\hat{R}(w_i), z - w_i\rangle$.

- **LHS.** Can be lower-bounded by

$$\max\left\{\inf_{i<t}\hat{R}(w_i), \hat{R}\left(\sum w_i/t\right)\right\}$$

Risk convergence

Lemma 7.2.

Suppose that $\hat{R}$ is convex, and let $G := \max_i \|g_i\|$. Let $\eta = 1/\sqrt{t}$. Then, for any z, we have

$$\frac{1}{t}\sum_{i<t} \hat{R}(w_i) \leq \hat{R}(z) + \frac{\|w_0 - z\|^2}{2\sqrt{t}} + \frac{G^2}{2\sqrt{t}} + \frac{1}{t}\sum_{i<t} \epsilon_i$$

where we use the shorthand $\epsilon_i = \langle g_i - \nabla \hat{R}(w_i), z - w_i \rangle$.

- **RHS.** Requires upper-bounding two quantities — will be discussed after the proof idea
  - $G := \max_i \|g_i\|$
  - $\frac{1}{t}\sum_{i<t} \epsilon_i$
    - Critically, ϵ_i may be dependent on ϵ_j

Proof idea

- **Proof idea.** Like GD, we can decompose the parameter updates:

$$\begin{aligned}\|w_{i+1} - z\|^2 &= \|w_i - z\|^2 - 2\eta\langle g_i, w_i - z\rangle + \eta^2\|g_i\|^2 \\ &= \|w_i - z\|^2 - 2\eta\langle \nabla \hat{R}(w_i), w_i - z\rangle + 2\eta\langle \nabla \hat{R}(w_i) - g_i, w_i - z\rangle + \eta^2\|g_i\|^2\end{aligned}$$

Add-and-Subtract; to exploit unbiasedness

Proof idea

- **Proof idea.** Like GD, we can decompose the parameter updates:

$$
\begin{aligned}
\|w_{i+1} - z\|^2 &= \|w_i - z\|^2 - 2\eta\langle g_i, w_i - z\rangle + \eta^2\|g_i\|^2 \\
&= \|w_i - z\|^2 - 2\eta\langle \nabla \hat{R}(w_i), w_i - z\rangle + 2\eta\langle \nabla \hat{R}(w_i) - g_i, w_i - z\rangle + \eta^2\|g_i\|^2 \\
&\leq \|w_i - z\|^2 - 2\eta(\hat{R}(w_i) - \hat{R}(z)) + 2\eta\langle \nabla \hat{R}(w_i) - g_i, w_i - z\rangle + \eta^2\|g_i\|^2
\end{aligned}
$$

Convexity

Proof idea

- **Proof idea.** Like GD, we can decompose the parameter updates:

$$
\begin{aligned}
\|w_{i+1} - z\|^2 &= \|w_i - z\|^2 - 2\eta\langle g_i, w_i - z\rangle + \eta^2\|g_i\|^2 \\
&= \|w_i - z\|^2 - 2\eta\langle \nabla \hat{R}(w_i), w_i - z\rangle + 2\eta\langle \nabla \hat{R}(w_i) - g_i, w_i - z\rangle + \eta^2\|g_i\|^2 \\
&\leq \|w_i - z\|^2 - 2\eta(\hat{R}(w_i) - \hat{R}(z)) + 2\eta\langle \nabla \hat{R}(w_i) - g_i, w_i - z\rangle + \eta^2\|g_i\|^2
\end{aligned}
$$

- Rearranging and scaling, we get the risk convergence

$$
\frac{1}{t}\sum_{i<t}\hat{R}(w_i) \leq \hat{R}(z) + \frac{\|w_0 - z\|^2 - \|w_t - z\|^2}{2\eta t} + \frac{1}{t}\sum_{i<t}\left(\epsilon_i + \frac{\eta}{2}\|g_i\|^2\right)
$$

  - We use the shorthand $\epsilon_i = \langle g_i - \nabla \hat{R}(w_i), z - w_i\rangle$

- Select the right η

Bounding the RHS

$$\frac{1}{t}\sum_{i<t}\hat{R}(w_i) \le \hat{R}(z) + \frac{\|w_0 - z\|^2}{2\sqrt{t}} + \frac{G^2}{2\sqrt{t}} + \frac{1}{t}\sum_{i<t}\epsilon_i$$

where we use the shorthand $\epsilon_i = \langle g_i - \nabla \hat{R}(w_i), z - w_i \rangle$.

- Now, back to bounding the quantities:
    - $G := \max_i \|g_i\|$
    - $\frac{1}{t}\sum_{i<t}\epsilon_i$
- We want to make sure that they diminish as $t \to \infty$

Side Note: Supremum of RVs

- Consider controlling the supremum

$$G := \max_i \|g_i\|$$

- **Simpler question.** Suppose that $X_1, \ldots, X_k \sim \mathcal{N}(0,1)$. Then, what is a nice UB on …? 🙋

$$\mathbb{E}[\max_i X_i]$$

Side Note: Supremum of RVs

- Consider controlling the supremum

$$G := \max_i \|g_i\|$$

- **Simpler question.** Suppose that $X_1, \ldots, X_k \sim \mathcal{N}(0,1)$. Then, what is a nice UB on ...? 🙋

$$\mathbb{E}[\max_i X_i]$$

- **Idea.**
  - First, note that

$$\max_i X_i = \log(\max_i \exp(X_i)) \leq \log(\sum \exp(X_i))$$

  - Then, take expectation to get:

$$\mathbb{E}[\max_i X_i] \leq \mathbb{E}\left[\log(\sum \exp(X_i))\right] \leq \log\left(\sum \mathbb{E}[\exp(X_i)]\right)$$

  - That is, at most of $\log k$

Controlling the gradient noise

- We further analyze

$$\epsilon_i = \langle g_i - \nabla \hat{R}(w_i), z - w_i \rangle$$

- We assume that we have the Martingale property, i.e.,

$$\mathbb{E}[g_i \mid w_{\leq i}] = \nabla \hat{R}(w_i)$$

- Then, we have a nice tool:

Theorem 7.8 (Azuma-Hoeffding).

Suppose that $(Z_i)_{i=1}^n$ is a Martingale difference sequence, i.e., $\mathbb{E}[Z_i \mid Z_{<i}] = 0$. Also, let $\mathbb{E}|Z_i| \leq R$. Then, with probability at least $1 - \delta$, we have

$$\sum_i Z_i \leq R\sqrt{2t \log(1/\delta)}$$

- Requires knowing the zero-mean-ness and UB on the mean absolute

Controlling the gradient noise

- Now, examine the case of $\epsilon_i = \langle g_i - \nabla \hat{R}(w_i), z - w_i \rangle$
- **Zero-mean.** We know that $\mathbb{E}[\epsilon_i \mid w_{\leq i}] = 0$.
- **UB on mean absolute.** We can proceed as:

$$\begin{aligned}\mathbb{E}|\epsilon_i| &= \mathbb{E}|\langle g_i - \nabla \hat{R}(w_i), w_i - z \rangle| \\ &\leq \mathbb{E}\|g_i - \nabla \hat{R}(w_i)\| \cdot \|w_i - z\| \\ &\leq \left(2 \cdot \text{gradient UB}\right) \cdot \left(\text{param radius}\right)\end{aligned}$$

Final form

- Summing up, we have

Lemma 7.3.

Let $\hat{R}$ be a convex function. Let G, D be uniform UBs on the gradients and parameter differences. Then, for $\eta = 1/\sqrt{t}$, the following holds with probability at least $1 - \delta$

$$\frac{1}{t}\sum_{i<t}\hat{R}(w_i) \leq R(z) + \frac{D^2}{2\sqrt{t}} + \frac{G^2}{2\sqrt{t}} + \frac{2GD\sqrt{2\log(1/\delta)}}{\sqrt{t}}$$

Next up

- NTK...