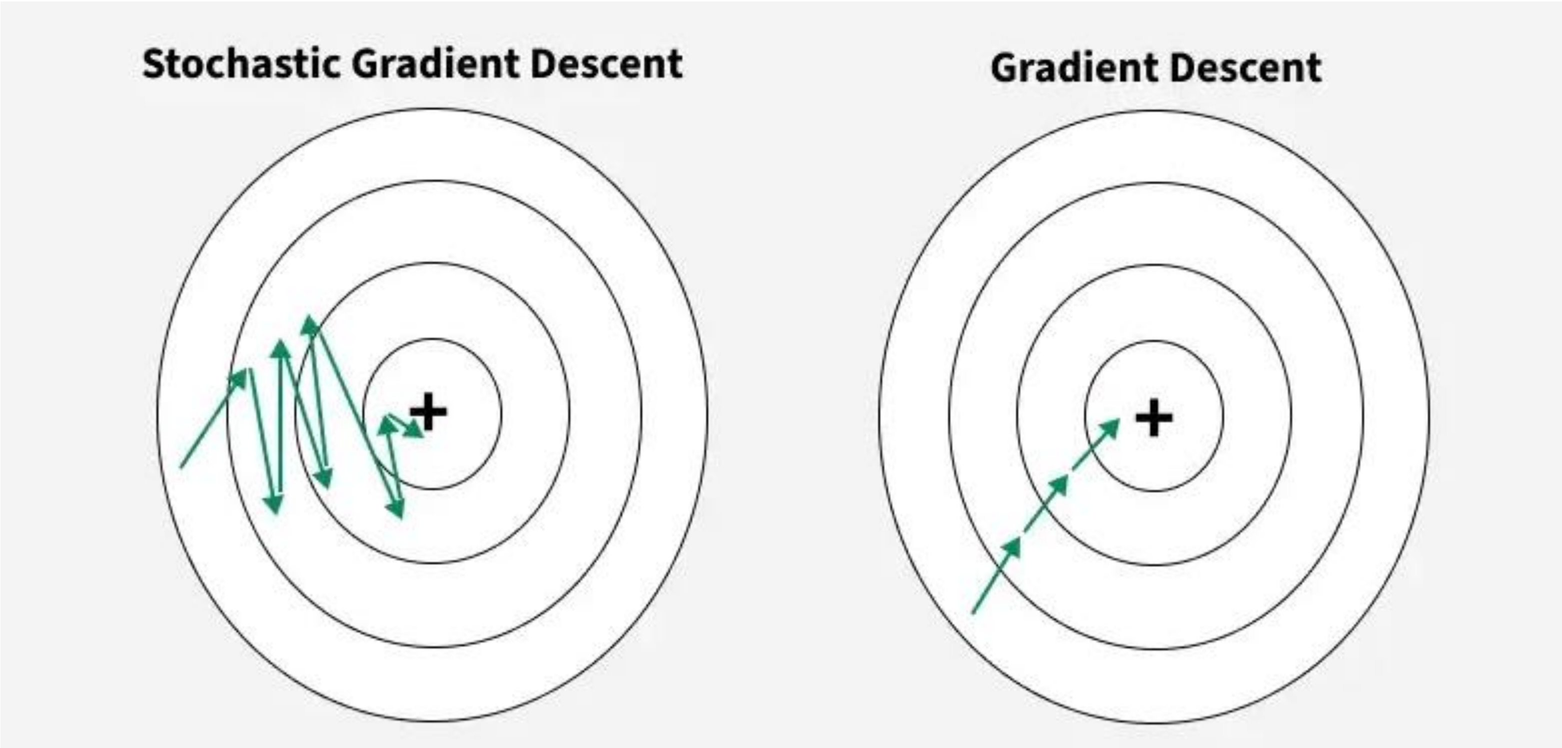
Rate of convergence of Stochastic Gradient Descent using Stein's method

Yuyang Wang, Felix Wang, Zedong Wang, Siva Theja Maguluri

Background

- Stochastic Gradient Descent (SGD) is an iterative method for optimizing an objective function
- Was introduced in the 1950
- Useful especially in high-dimensions which reduces the high computational burden
- Today, mainly used as an optimization tool in Machine Learning

SGD



More Formally

- Stochastic Gradient Descent defined by:

$$X_{k+1}^{(\alpha)} = X_k^{(\alpha)} + \alpha(-\nabla f(X_k^{(\alpha)}) + w_k)$$

- Using the Scaled Iterate, defined below, we can find convergence to the limit

$$Y_k^{(\alpha)} = \frac{X_k^{(\alpha)} - x^*}{\sqrt{\alpha}}$$

Convergence

$$Y_k^{(\alpha)} \xrightarrow{k\to\infty} Y^\alpha \quad \text{(Stationary Distribution)}$$

$$Y^\alpha \xrightarrow{\alpha\to 0} \mathcal{N}(0, \Sigma)$$

Problem Setup

- We know that as α goes to 0, Y goes to Gaussian

- What is the rate of convergence?
  - Distance between distributions

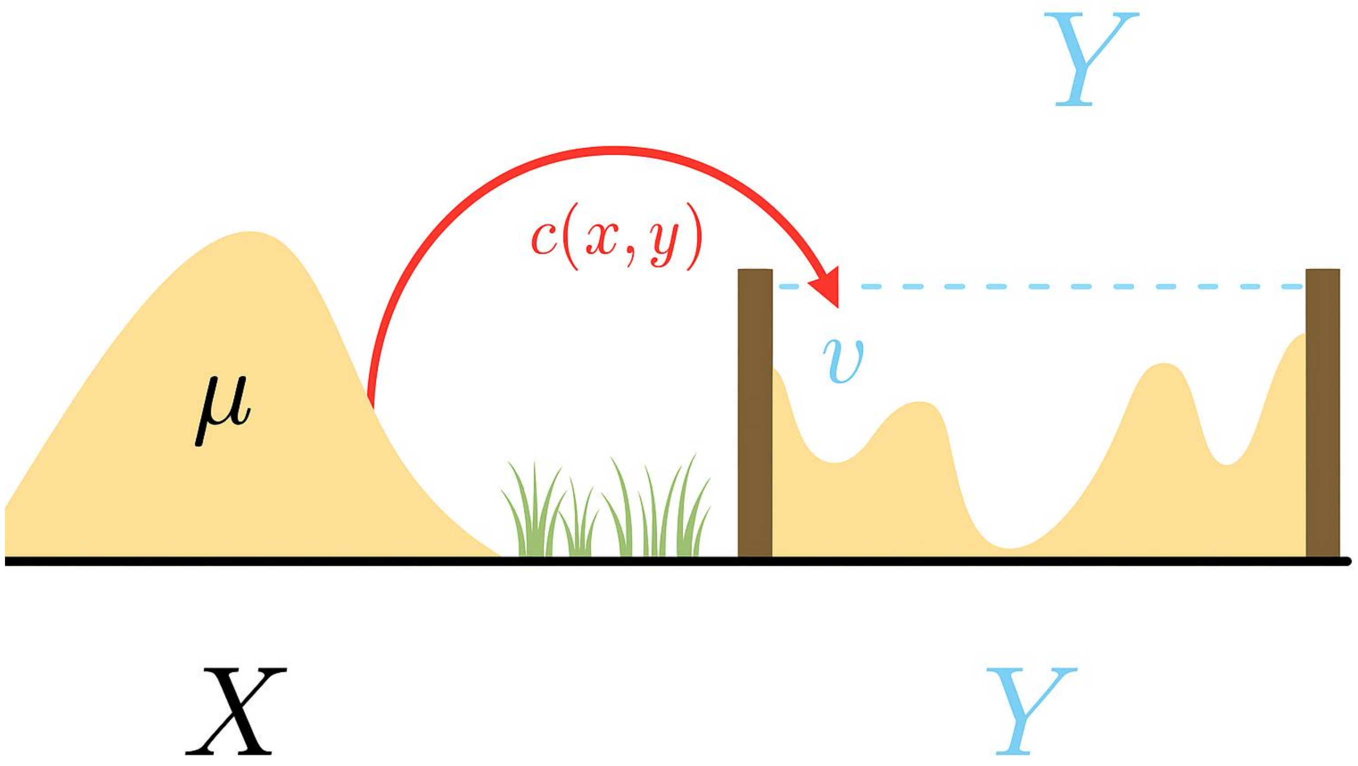
- Why important
  - Other examples CLT

Wasserstein Distance

$$d_W(W, Z) = \sup_{h \in H} |\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]|$$

$$H = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y) \leq |x - y|\}$$

Wasserstein Distance



Illuminative example: f(x) = $x^2/2$

- The new Stochastic Gradient Descent would become:

$$X_{k+1}^{(\alpha)} = (1-\alpha)X_k^{(\alpha)} + \alpha w_k$$

- With $x^* = 0$, the new Scaled iterate is:

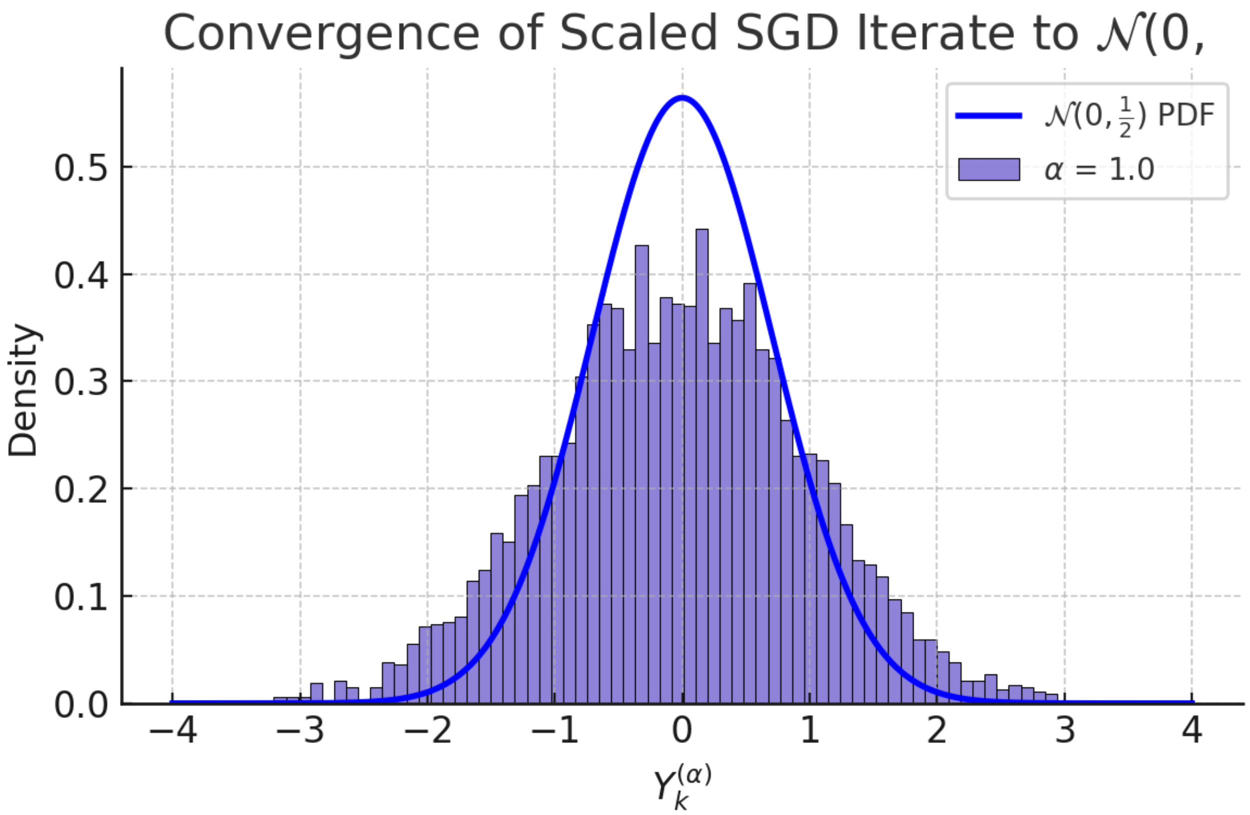
$$Y_k^{(\alpha)} = \frac{X_k^{(\alpha)}}{\sqrt{\alpha}}$$

Illuminative example: $f(x) = x^2/2$

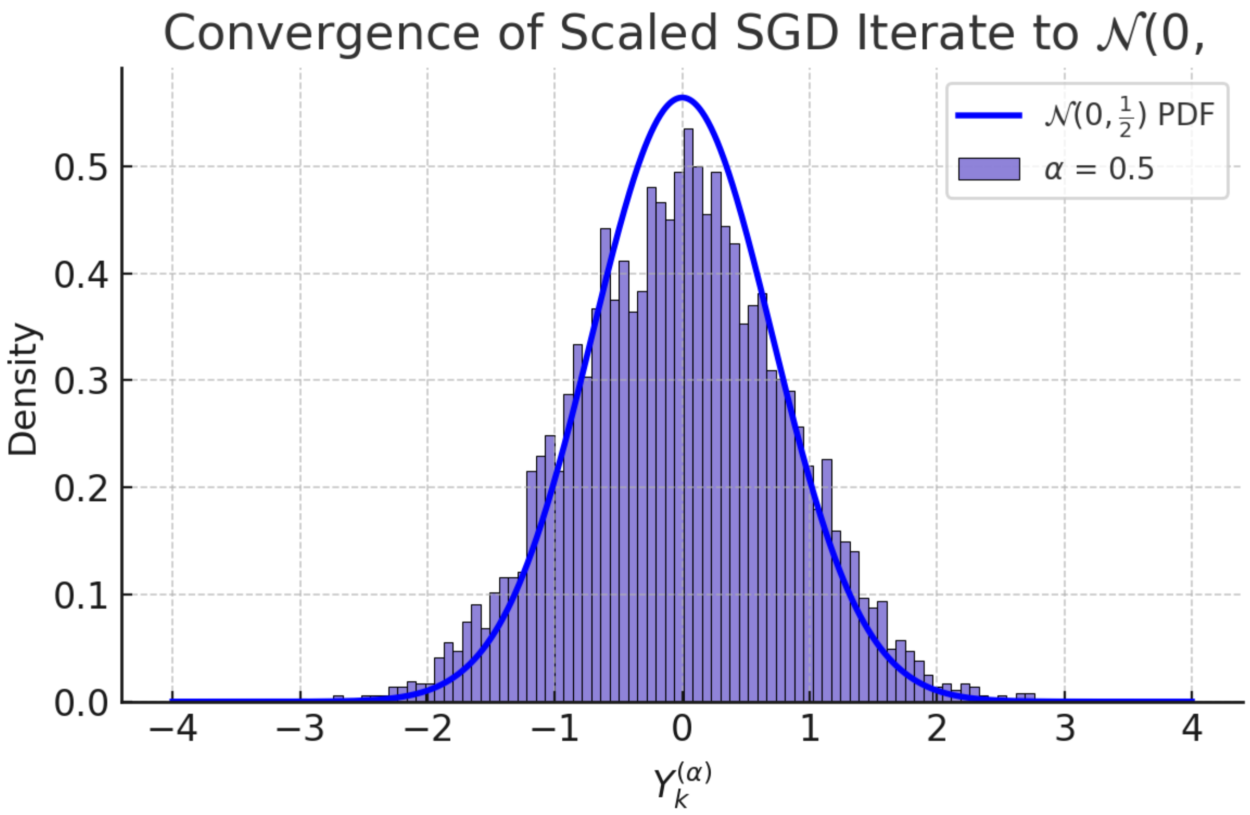
- Based off of Zaiwei's paper [1], they found that:

$$Y_k^{(\alpha)} \to \mathcal{N}(0, \frac{1}{2-\alpha})$$

To Show This:

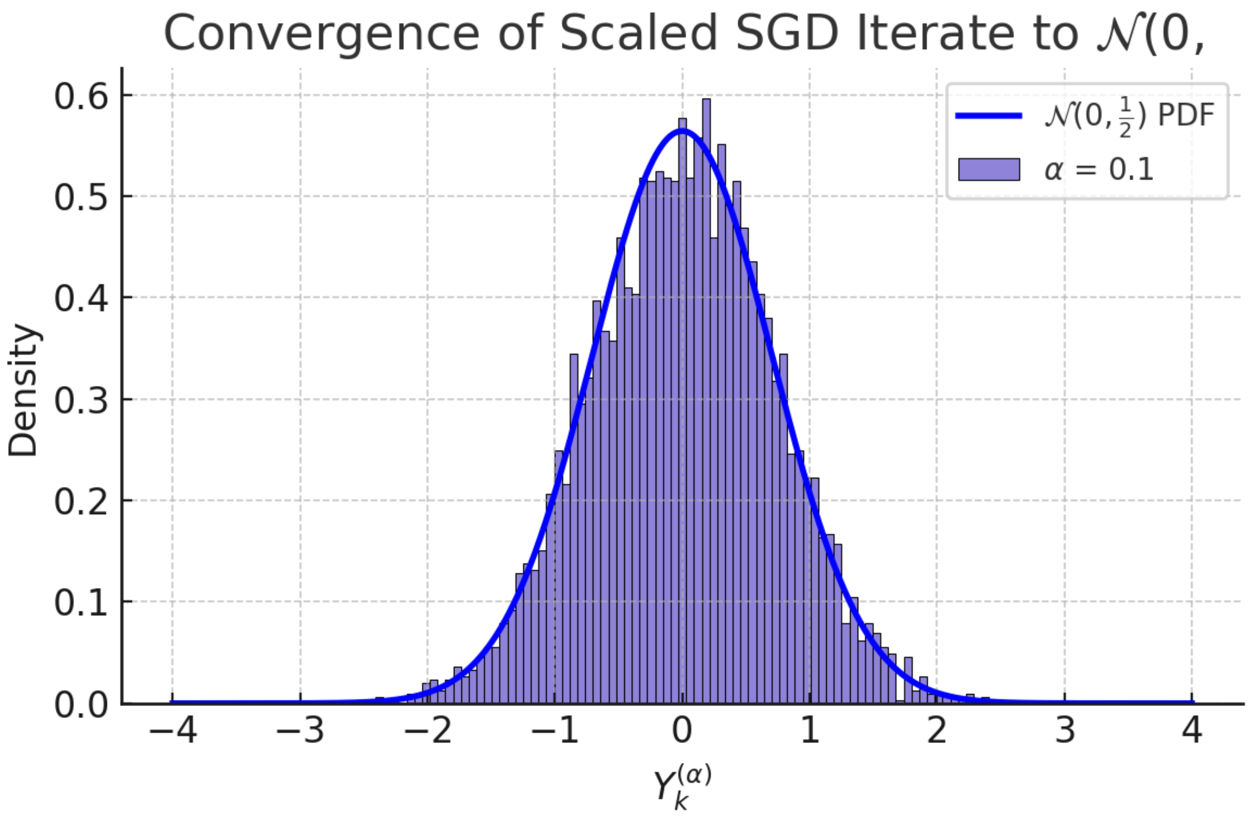

Convergence of Scaled SGD Iterate to $\mathcal{N}(0,$
$\mathcal{N}(0, \frac{1}{2})$ PDF
$\alpha = 1.0$
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−3
−2
−1
0
1
2
3
4
$Y_k^{(\alpha)}$

To Show This:



Convergence of Scaled SGD Iterate to $\mathcal{N}(0,$
$\mathcal{N}(0, \frac{1}{2})$ PDF
$\alpha = 0.5$
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−3
−2
−1
0
1
2
3
4
$Y_k^{(\alpha)}$

To Show This:



Convergence of Scaled SGD Iterate to $\mathcal{N}(0,$
$\mathcal{N}(0, \frac{1}{2})$ PDF
$\alpha = 0.1$
Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−4
−3
−2
−1
0
1
2
3
4
$Y_k^{(\alpha)}$

Result:

$$d_W(W, Z) \le O(\sqrt{\alpha})$$

$$d_{\mathrm{W}}(W,Z) \;\le\; \left(\sqrt{2\pi(2-\alpha)\mathbb{E}[|w_i|^4]} + \frac{8(2-\alpha)^{\frac{3}{2}}}{3}\mathbb{E}[|w_i|^3]\right)\alpha^{\frac{1}{2}}$$

**third and fourth moments of w_i exists*

- Idea of Proof:
- Step 1: Build the Stein pair (W,W')

$$W = \frac{Y_k^{(\alpha)} - \mathbb{E}[Y_k^{(\alpha)}]}{\sqrt{\mathrm{Var}(Y^{(\alpha)})}} \qquad W' = W - \frac{1}{\sigma}(1-\alpha)^{k-1-i}\sqrt{\alpha}\, w_i + \frac{1}{\sigma}(1-\alpha)^{k-1-i}\sqrt{\alpha}\, w_i'.$$

$$= \frac{1}{\sigma}\sum_{i=0}^{k-1}(1-\alpha)^{k-1-i}\sqrt{\alpha}\, w_i$$

- Step 2 (From [2]):

If (W, W') is an a-Stein pair with $\mathbb{E}[W^2] = 1$ and $Z \sim \mathcal{N}(0,1)$, then

$$d_{\mathrm{W}}(W, Z) \le \frac{\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2 \mid W])}}{\sqrt{2\pi}\, a} + \frac{\mathbb{E}|W'-W|^3}{3a}.$$

- Step 3: Doing computations

- LIMIT: Only works if we can solve the iteration.

General case

Assumption 1. The noise sequences $\{w_k\}$ is independent and identically distributed with mean zero and a positive definite covariance $\Sigma \in \mathbb{R}^{d\times d}$.

Definition 2. A differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and σ-convex with respect with $\|\cdot\|_2$ if and only if

$$h(y) \leq h(x) + \langle \nabla h(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2, \qquad \text{(L-smooth)}$$

$$h(y) \geq h(x) + \langle \nabla h(x), y - x \rangle + \frac{\sigma}{2}\|x - y\|_2^2, \qquad (\sigma\text{-convex})$$

for all $x, y \in \mathbb{R}^d$.

Assumption 2. The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is second differentiable and is both L-smooth and σ-convex.

Assumption 3. The objective function is thrice differentiable and $\sup ||f_{ijk}||_\infty = M < \infty$ for some $M \in \mathbb{R}$, which means all it's third derivatives are uniformly bounded.

Result

Under these assumptions, the following holds:

$$d_W(Y^{(\alpha)}, Y) \leq L_1\sqrt{\alpha} + L_2\alpha$$

such that

$$L_1 = d^3 CM \frac{\mathrm{Trace}(\Sigma)}{\sigma} + d^3 C (\sum_{ij}^{d} |\Sigma_{ij}|) \,\mathrm{Trace}(\Sigma)^{\frac{1}{2}}$$

and

$$L_2 = dCL^2 \frac{\mathrm{Trace}(\Sigma)}{\sigma} + d^3 C (\sum_{ij}^{d} |\Sigma_{ij}|) (\frac{\mathrm{Trace}(\Sigma)}{\sigma})^{\frac{1}{2}} \,,$$

where M and C are independent from α.

ALSO, the uniqness conjecture in Zaiwei's paper is solved using characteristic method.

Stein's Method

Goal: compare Y to Z~N(0,1) (e.g., in Wasserstein distance).

Stein operator (normal): $Lf(x) = f'(x) - xf(x)$ $\quad E[Lf(Z)] = 0$

Stein equation for a test function h: $Lg_h(y) = h(y) - E[h(Z)]$

Solving this gives us a larger class of test functions: $g_h(Y)$

$$\begin{aligned} d_W(Y,Z) &= sup_{h \in Lip\{1\}}\{\mathbb{E}[h(Y) - h(Z)]\} \\ &\leq sup_{g_h \in \mathbf{F}}\{\mathbb{E}[Ag_h(Y)]\} \end{aligned}$$

Also works for higher dimension and other target distributions

Idea of Proof

Step 1: Construct Stein operator via exchangable pairs.

Proposition 3. Let X and X' be an exchangeable pair. Considering the operator

$$Af(x) := \mathbb{E}[f(X') - f(X)|X = x] \, .$$

Then

$$\mathbb{E}[Af(X)] = 0$$

for all f integrable.

Step 2: Changing the distance between two random variables into the difference of two Stein operators.

$$\begin{aligned} d_W(Y^{(\alpha)}, Y) &= sup_{h \in Lip\{1\}}\{\mathbb{E}[h(Y^{(\alpha)}) - h(Y)]\} \\ &\leq sup_{g_h \in \mathbf{F}}\{\mathbb{E}[Ag_h(Y^{(\alpha)})]\} \\ &= sup_{g_h \in \mathbf{F}}\{\mathbb{E}[Ag_h(Y^{(\alpha)}) - A^{(\alpha)}g_h(Y^{(\alpha)})]\} \end{aligned}$$

Step 3: Using Taylor expension to estimate the difference.

Future Work

- The general rescaling factor
  - Give a reasonable guess

- Contractive
  - Linear

Thank You

References

- [1] Zaiwei Chen, Shancong Mou, and Siva Theja Maguluri. Stationary behavior of constant stepsize sgd type algorithms: An asymptotic characterization. Proc. ACM Meas. Anal. Comput. Syst., 6(1), February 2022.
- [2] Nathan Ross. Fundamentals of stein’s method. Probability Surveys, 8:210–293, 2011.