SGD - Rate of convergence to Gaussian using Steins method

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

Industrial and Systems Engineering, Georgia Institute of Technology

Georgia Tech College of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

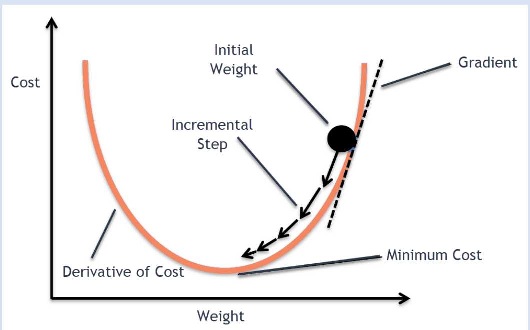
Illuminative example $f(x) = x^2/2$:

- **Stochastic Gradient Descent:**

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

- **Scaled Iterate**

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



- **Stochastic Gradient Descent:**

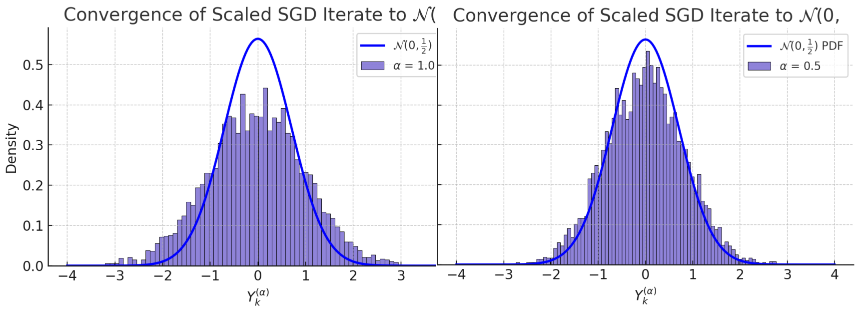
$$X_{k+1}^{(\alpha)} = (1-\alpha)X_k^{(\alpha)} + \alpha w_k$$

- **Scaled Iterate:**

$$Y_k^{(\alpha)} = \frac{X_k^{(\alpha)}}{\sqrt{\alpha}}, Y_k^{(\alpha)} \to \mathcal{N}(0, \frac{1}{2-a}) \quad [1]$$

- **Updated SGD:**

$$Y_k^{(\alpha)} = (1-\alpha)^k Y_0^{(\alpha)} + \sum_{i=0}^{k-1} (1-\alpha)^{k-1-i}\sqrt{\alpha}\, w_i.$$



Results:

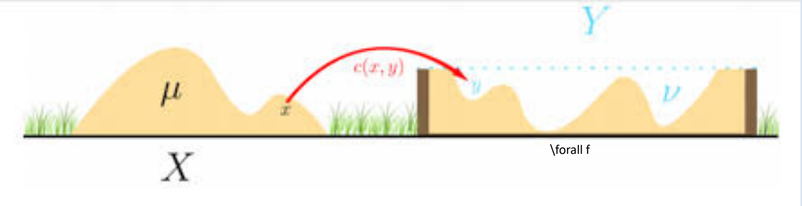
$$d_W(W,Z) = \lim_{k\to\infty} d_W(W_k, Z) = O(\sqrt{\alpha})$$

General 1 Dimensions $f: \mathbb{R} \to \mathbb{R}$

Wasserstein Distance:

$d_W(W,Z) = \sup_{h\in H} |\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]|, \quad H = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y) \leq |x-y|\}$

A way to think about the is the optimal transport problem:



Assumptions:

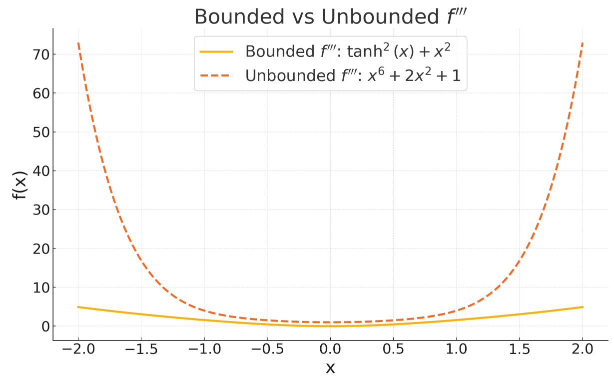
1. f(x) is thrice differtiable such that $\boxed{||f'''(x)||_\infty = C < \infty}$
2. L-smooth
3. σ-convex
4. $\mathbb{E}[w_k] = 0, \mathrm{Var}[w_k] = \Sigma$

Theorem:

1. There exists $a' > 0$, $\forall \alpha \in (0, a')$, then $Y_k^{(\alpha)} \xrightarrow{d} Y^{(\alpha)}$
2. For $\{\alpha_k\} \in (0, a')$, $\lim_{k\to\infty} \alpha_k = 0$, then

$$Y^{(\alpha)} \xrightarrow{d} Y \sim \mathcal{N}(0, \Sigma_y), \quad \Sigma_y = \frac{\Sigma}{2f''(x^*)}$$

3. $d_W(Y^{(\alpha)}, Y) \leq L_1\sqrt{\alpha} + L_2\alpha = \boxed{O(\sqrt{\alpha})}$



Proof for $f(x) = x^2/2$

Steins Method:

$$\forall f, \quad \mathbb{E}[A f(X)] = 0 \leftrightarrow X \sim Z$$

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

Definitions:

$$W_k = \frac{Y_k^{(\alpha)} - \mathbb{E}[Y_k^{(\alpha)}]}{\sqrt{\mathrm{Var}(Y_k^{(\alpha)})}}$$

$$W_k' = W_k - \frac{1}{\sigma}(1-\alpha)^{k-1-i}\sqrt{\alpha}\, w_i + \frac{1}{\sigma}(1-\alpha)^{k-1-i}\sqrt{\alpha}\, w_i'$$

$$\sigma^2 = \mathrm{Var}[Y_k^{(\alpha)}]$$

Exchangeable pairs: $(W, W') \overset{d}{=} (W', W)$

a-Stein pair: $E[W \mid W'] = (1-a)W$

Proposition 1 [2]:
Suppose (W,W') is a-Stein pair, $\mathbb{E}[W^2] = 1$, Z is standard normal, then

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

Proof Sketch

1. (W_k, W_k') form a (1/n)-Stein pair:
2. Proposition 1: $\lim_{k\to\infty} d_W(W_k, Z) = L_1\sqrt{\alpha}, \quad L_1 > 0$

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.