CIS 505: Optimization Project

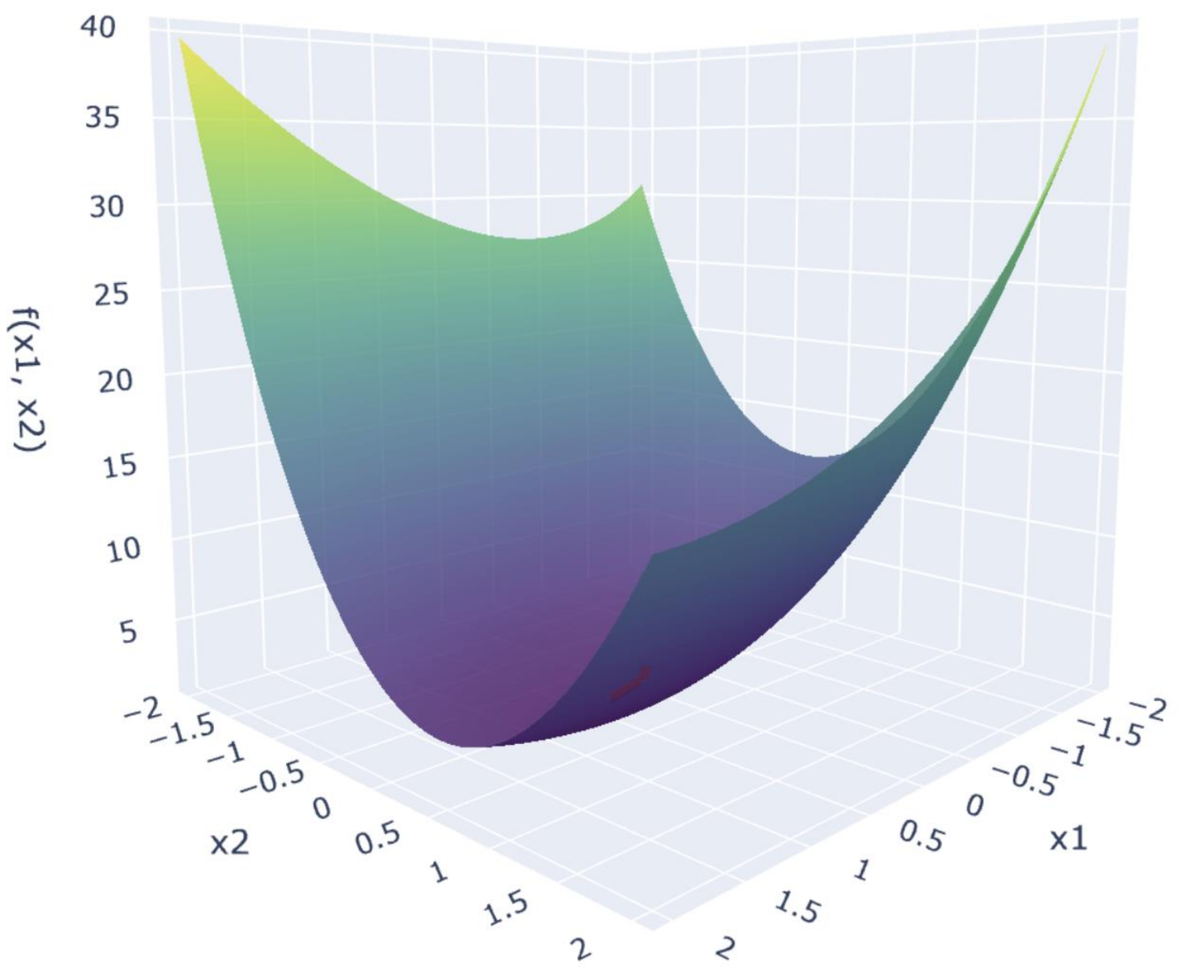
Adam & Nadam:

A comparative study against Steepest Descent Algorithm

Meeshawn Marathe

UMID: 4575 4188

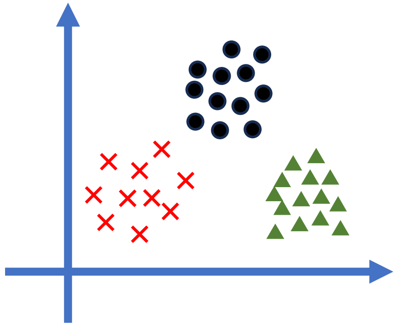
Nov 14th, 2023



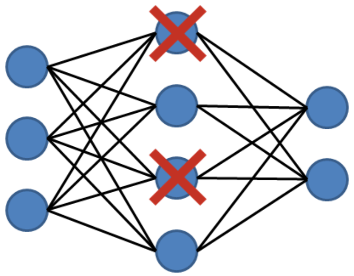
Adam/Nadam - Motivation

- Stochastic Objective Functions

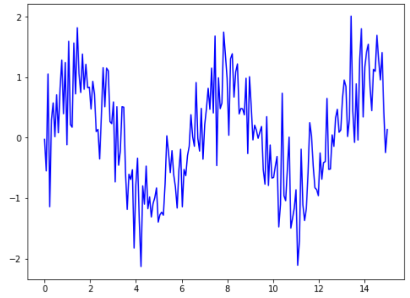
- Data Subsampling



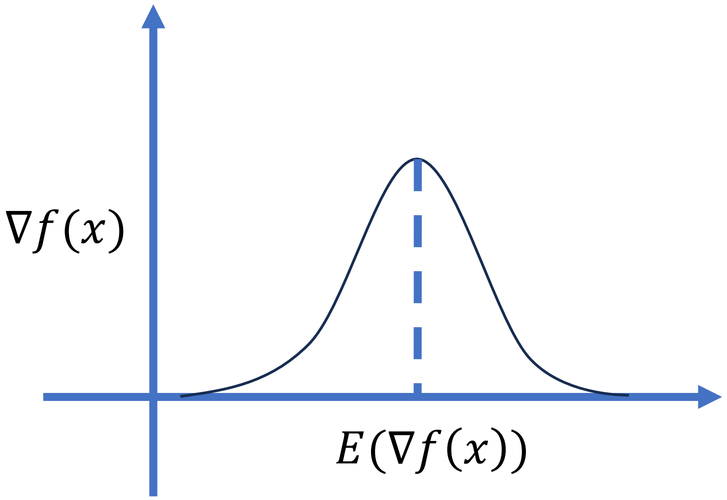
- Drop out Regularization



- Noisy dataset



Noisy Gradients



Adam - Introduction

- Efficient 1st order method Stochastic Optimization method
- ADAM: **Ada**ptive **M**oment Estimation
- Adaptive learning rates → 1st order moment ($\mathrm{E}[\nabla f(x)]$, mean) & 2nd order moment ($\mathrm{E}[(\nabla f(x)^2]$, uncentered variance)
- Exponential Moving Averaging → 1st order and 2nd order momentums

$$\underbrace{\underbrace{\ldots\ldots\nabla \mathrm{f}(\mathrm{x}_{i-2}), \nabla \mathrm{f}(\mathrm{x}_{i-1})}_{m_{i-1}, v_{i-1}}, \nabla \mathrm{f}(\mathrm{x}_{i})}_{m_i, v_i}, \nabla \mathrm{f}(\mathrm{x}_{i+1})\ldots\ldots$$

....

Adam – Algorithm[1]

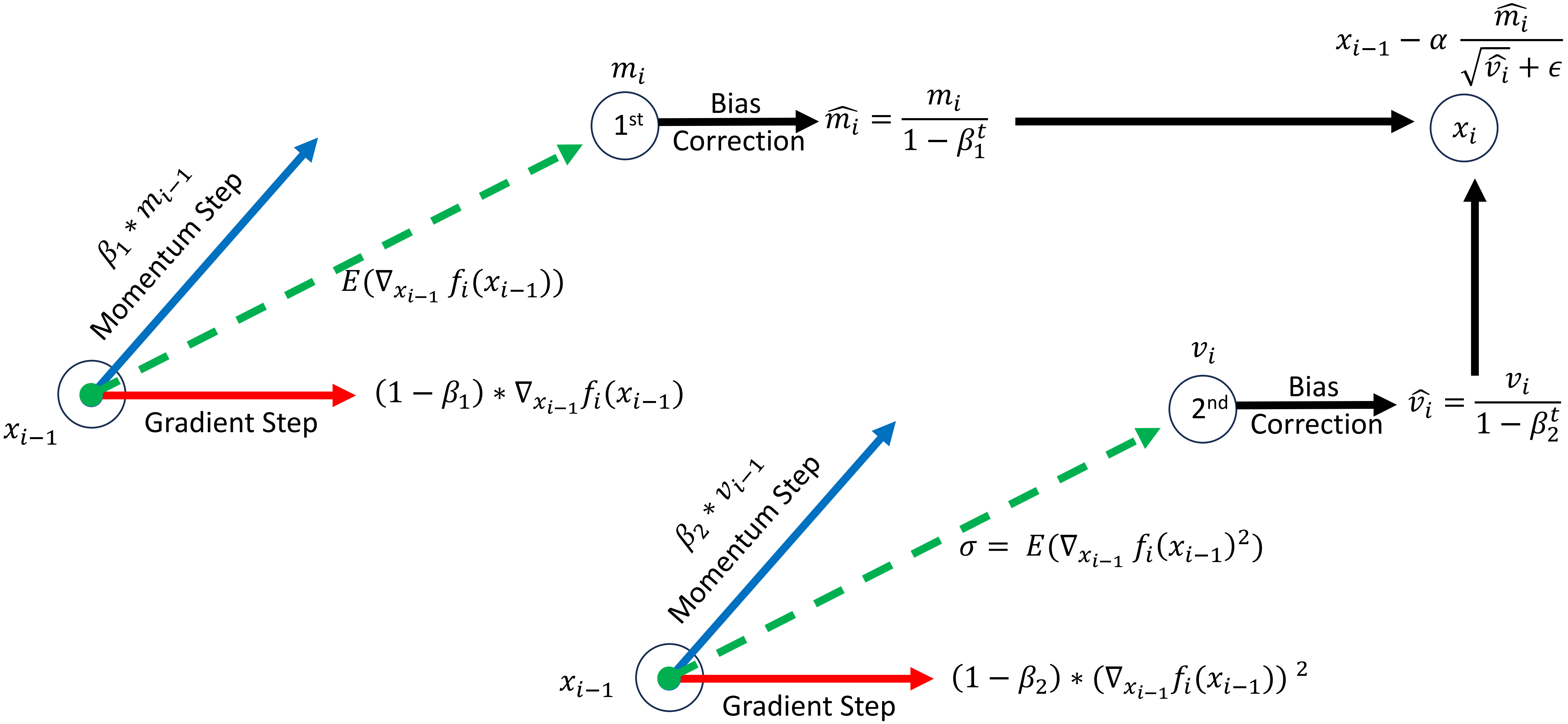
Algorithm 1 Adaptive Moment Estimation

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector

Hyperparameters

$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
  $t \leftarrow t + 1$
  $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
  $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
  $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
  $\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
  $\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
  $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

Adam - Algorithm



Adam - Algorithm

$$x_{i-1} - \alpha \frac{\widehat{m_i}}{\sqrt{\widehat{v_i}} + \epsilon} \quad \leftrightarrow \quad \alpha_t = \alpha \cdot \sqrt{1-\beta_2^t}/(1-\beta_1^t) \text{ and } \theta_t \leftarrow \theta_{t-1} - \alpha_t \cdot m_t/(\sqrt{v_t} + \hat{\epsilon})$$

x_i

Adam – Pros/Cons Over Steepest Descent

- Faster Convergence compared to Steepest Descent.
- Works well with noisy (stochastic objectives) and sparse gradients.
- Naturally performs stepsize annealing (Adaptive stepsizing)
- Stepsize bounded approximately by "*stepsize*" hyperparameter
- Parameter update values invariant to rescaling of gradient (ratio of moments)
- Scales well to large scale, high-dimensional ML problems
- Requires more space compared to Steepest Descent to store the n-dimensional moments.
- Is computationaly expensive compared to Steepest Descent due to the additional computations of the expected moments.

Adam – Pros/Cons

- Using the previous gradients instead of the previous updates allows the algorithm to continue changing direction even when the learning rate has annealed significantly toward the end of training, resulting in more precise fine-grained convergence.
- It also allows the algorithm to straightforwardly correct for the initialization bias that arises from initializing the momentum vector to 0

Nadam – Introduction

- Nadam: **N**esterov Accelerated Gradient + **Adam**
- Adam: Adaptive Learning Rate + Momentum
- Regular momentum can be shown conceptually and empirically to be inferior to a similar algorithm known as Nesterov's accelerated gradient (NAG).

Nadam – Algorithm[2]

Algorithm 3 Nesterov-accelerated Adaptive Moment Estimation (Nadam)

Require: $\alpha_0, \ldots, \alpha_T; \mu_0, \ldots, \mu_T; \nu; \epsilon$: Hyperparameters

$\mathbf{m}_0; \mathbf{n}_0 \leftarrow 0$ (first/second moment vectors)

while θ_t not converged **do**

$\mathbf{g}_t \leftarrow \nabla_{\theta_{t-1}} f_t(\theta_{t-1})$

$\mathbf{m}_t \leftarrow \mu_t \mathbf{m}_{t-1} + (1 - \mu_t)\mathbf{g}_t$

$\mathbf{n}_t \leftarrow \nu \mathbf{n}_{t-1} + (1 - \nu)\mathbf{g}_t^2$

$\hat{\mathbf{m}} \leftarrow (\mu_{t+1}\mathbf{m}_t/(1 - \prod_{i=1}^{t+1} \mu_i)) + ((1 - \mu_t)\mathbf{g}_t/(1 - \prod_{i=1}^{t} \mu_i))$

$\hat{\mathbf{n}} \leftarrow \nu \mathbf{n}_t/(1 - \nu^t)$

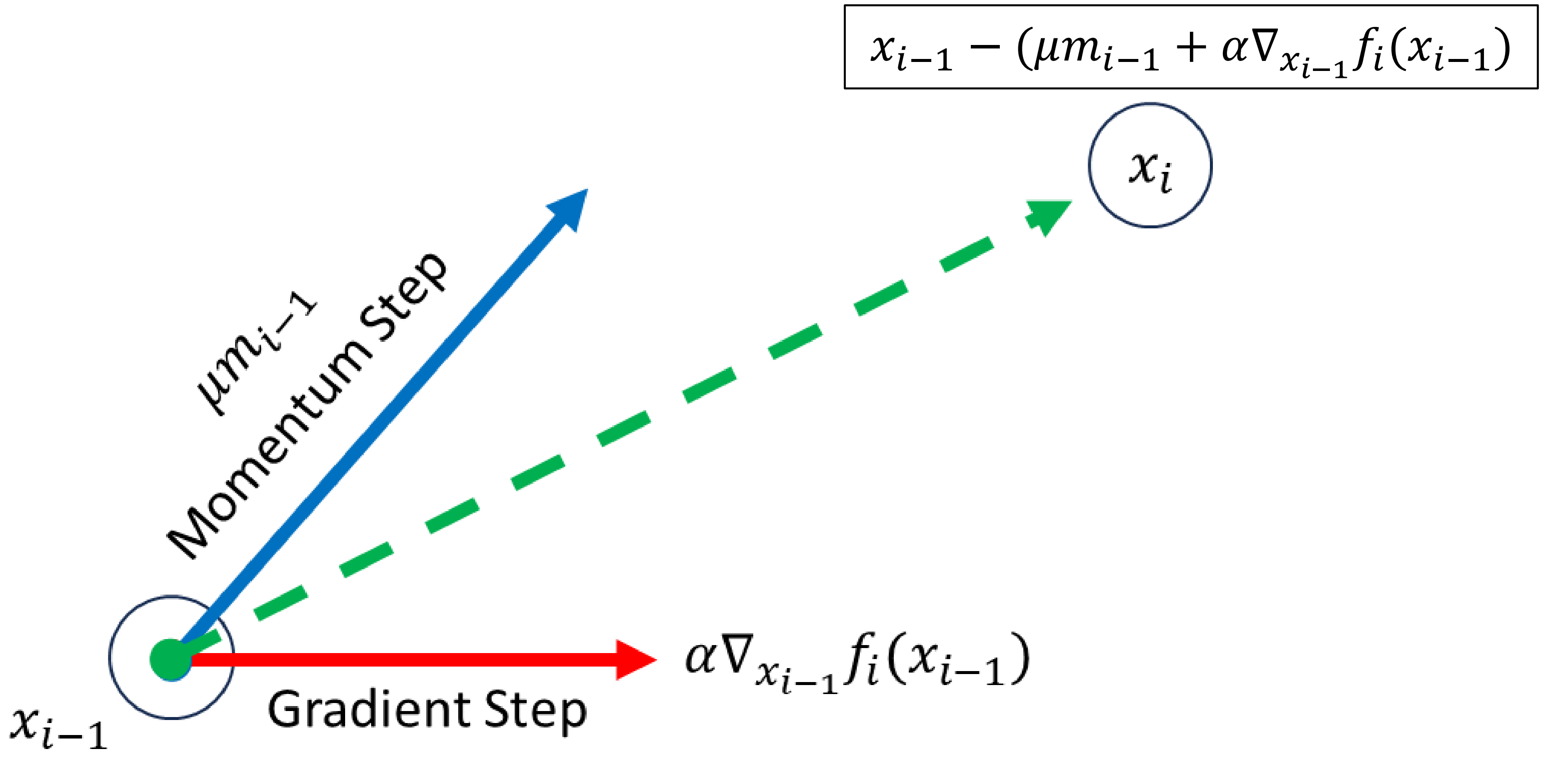
$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha_t}{\sqrt{\hat{\mathbf{n}}_t + \epsilon}} \hat{\mathbf{m}}_t$

end while

return θ_t

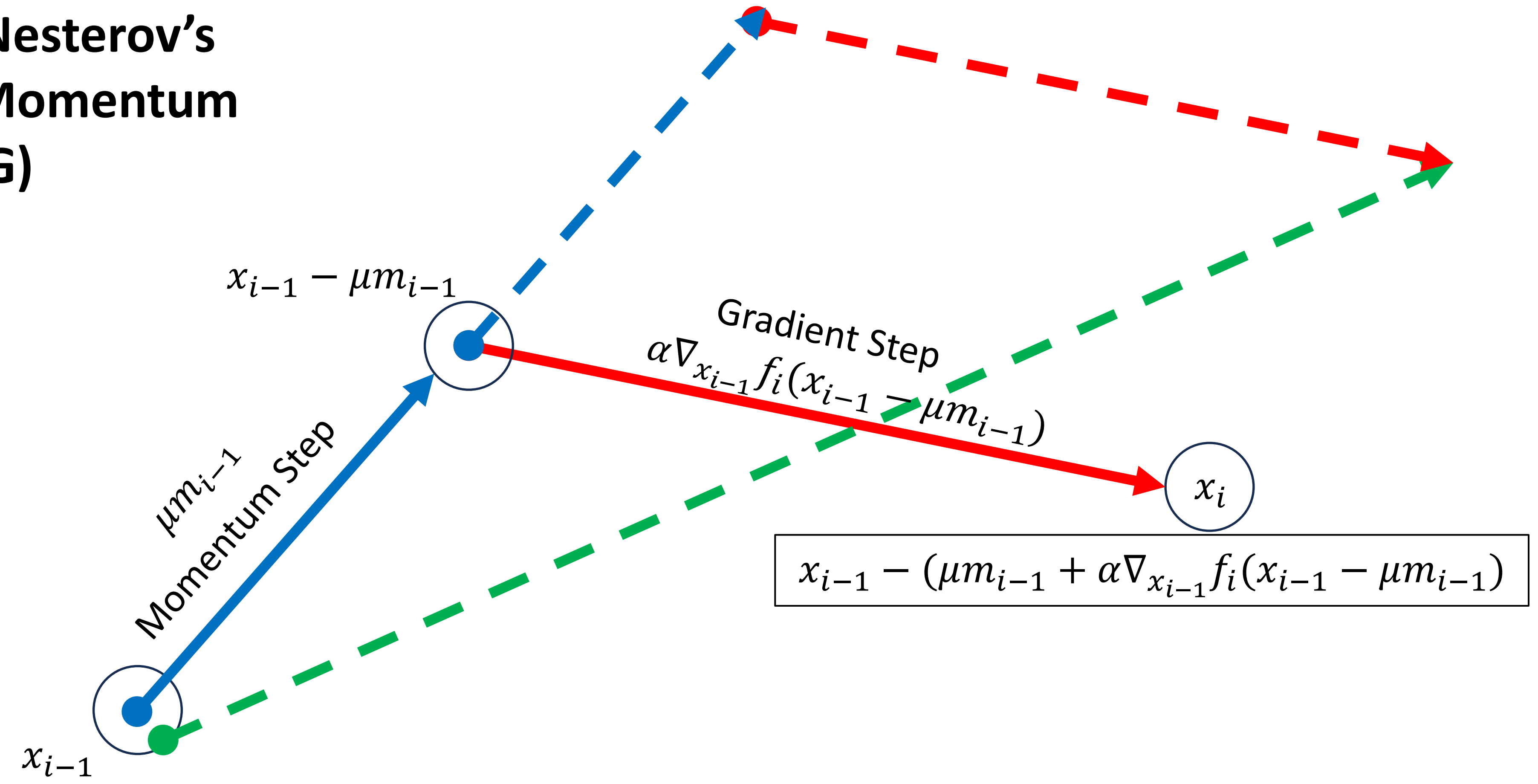
Nadam – Introduction

Regular Momentum



Nadam – Introduction

Inspiration: **Nesterov's Accelerated Momentum (NAG)**



Nadam – Pros/Cons

- Same Pros as Adam
- Is quicker to converge than Adam due to the additional acceleration provided by NAG
- Is slightly computationally expensive compared to Adam due to additional gradient computation in the momentum step.
- Requires more space compared to Steepest Descent to store the n-dimensional moments.
- Is computationaly expensive compared to Steepest Descent due to the additional computations of the expected moments.

Steepest Descent - Algorithm

Algorithm 2 Steepest Descent

Require: α: The fixed learning rate
Require: $f_i(\theta)$: Stochastic objective function parameterized by θ and indexed by timestep i
Require: θ_0: The initial parameters
  while θ_t not converged **do**
    $t \leftarrow t + 1$
    $\mathbf{g}_t \leftarrow \nabla_{\theta_{t-1}} f_t(\theta_{t-1})$
    $\theta_t \leftarrow \theta_{t-1} - \alpha \mathbf{g}_t$
  end while
  return θ_t

Experimental Setup

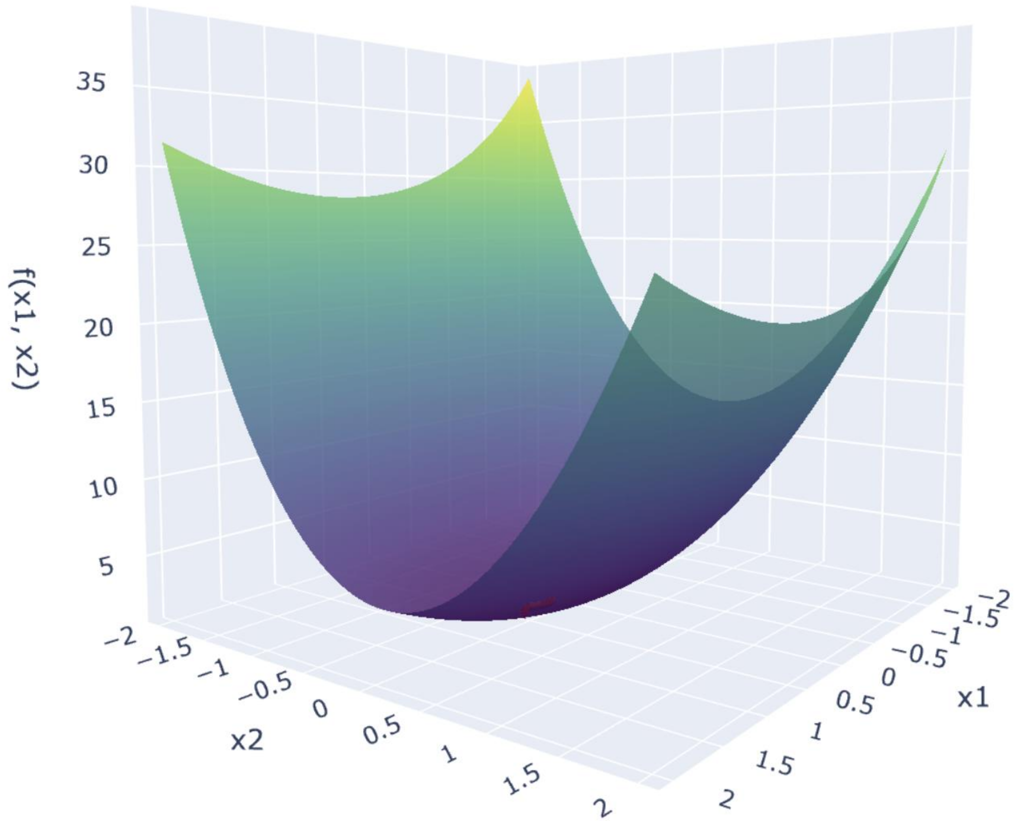
$$f(x_1, x_2) = e^{-x_1} + e^{-x_2} + x_1^2 + 5x_2^2$$

$$X_0 = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Hyperparameters

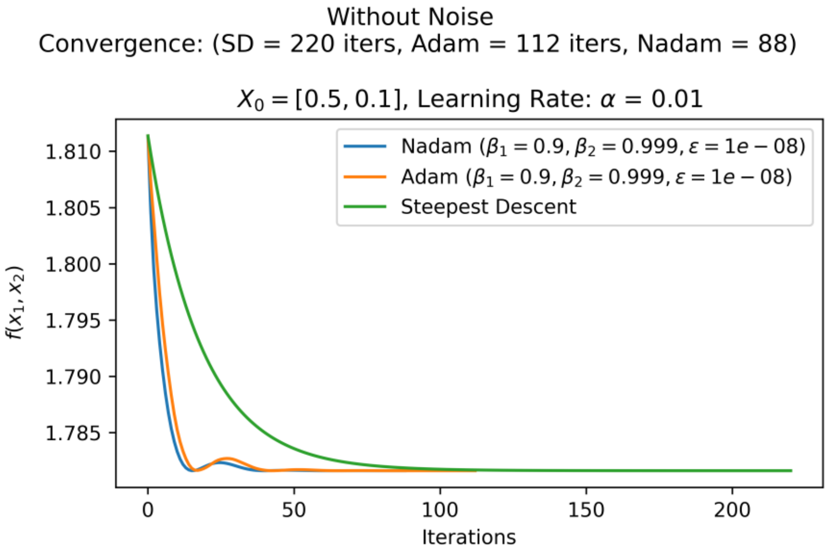
$$\beta_1, \beta_2 = (0.9, 0.999), (0.9, 0.5), (0.9,0$$

$$\alpha_0 = (0.001, 0.01)$$

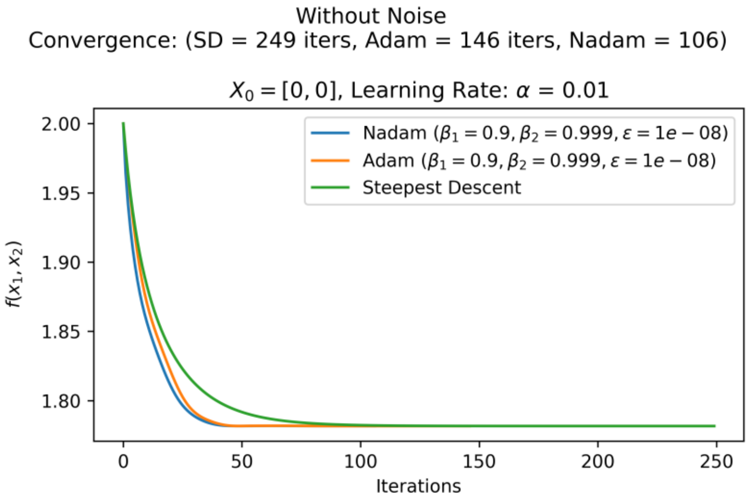


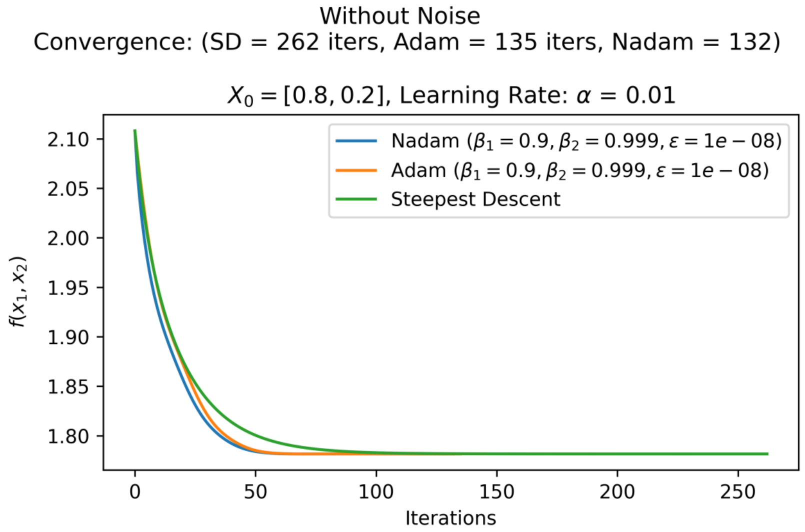
Numerical Results

Convergence Criterion: $x_i - x_{i-1} = 1e - 5$



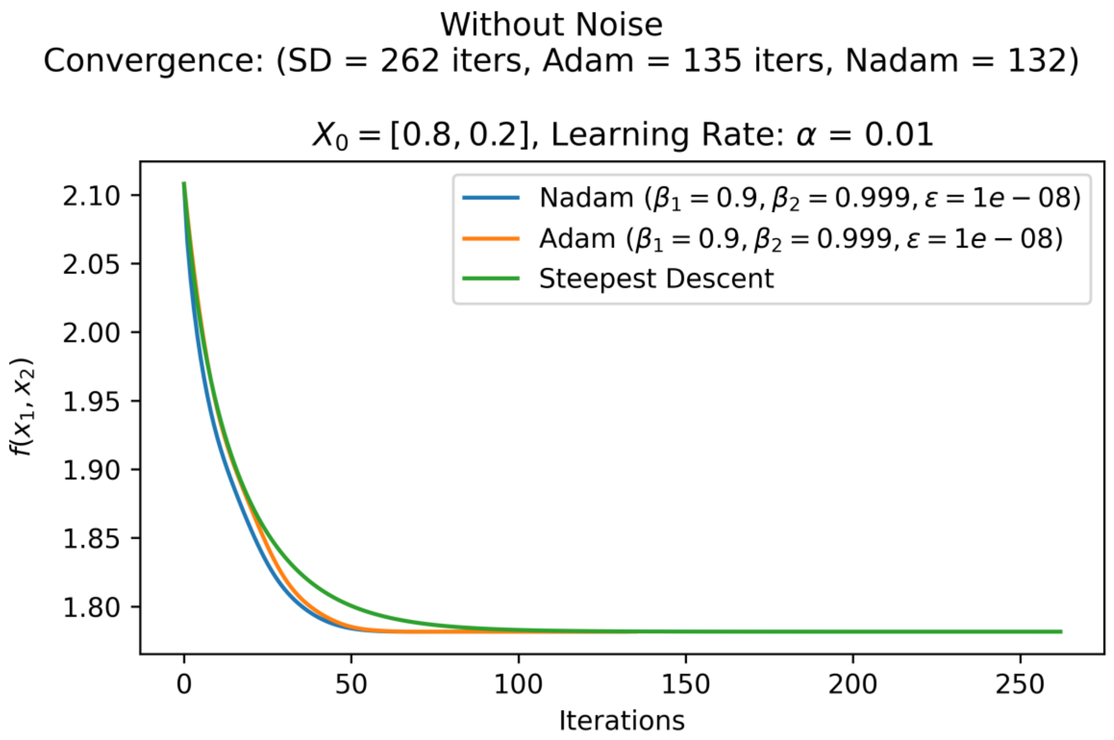
Convergence at different initial starting values

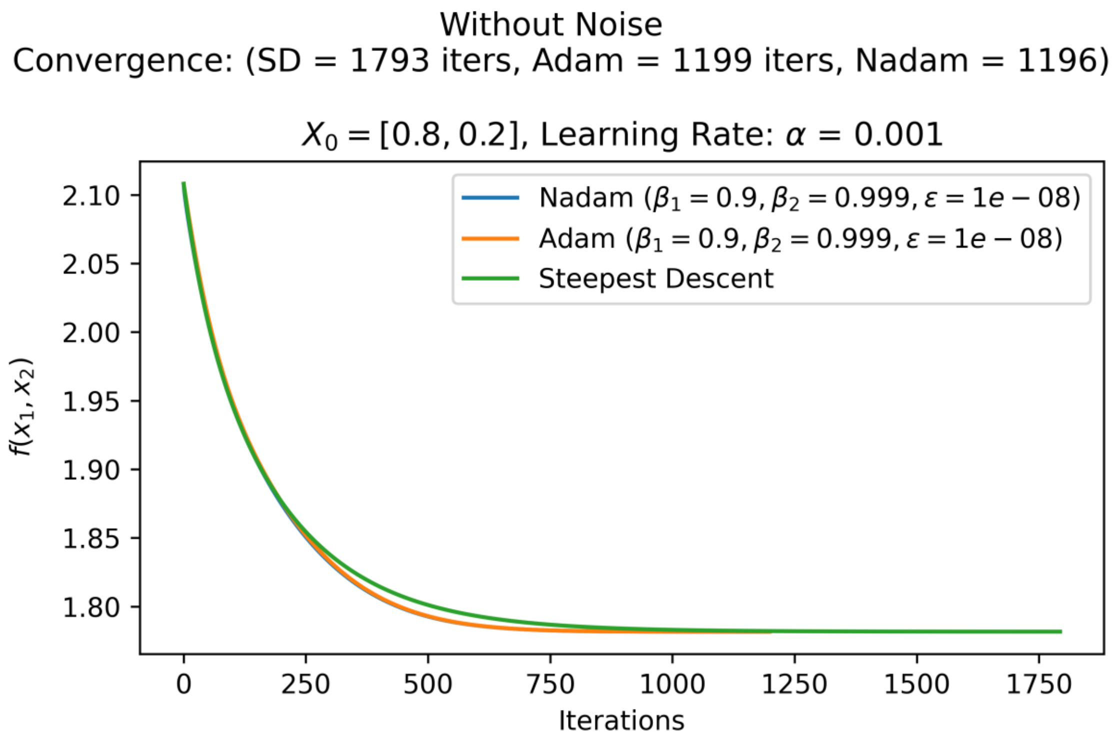




Numerical Results

Convergence Criterion: $x_i - x_{i-1} = 1e - 5$

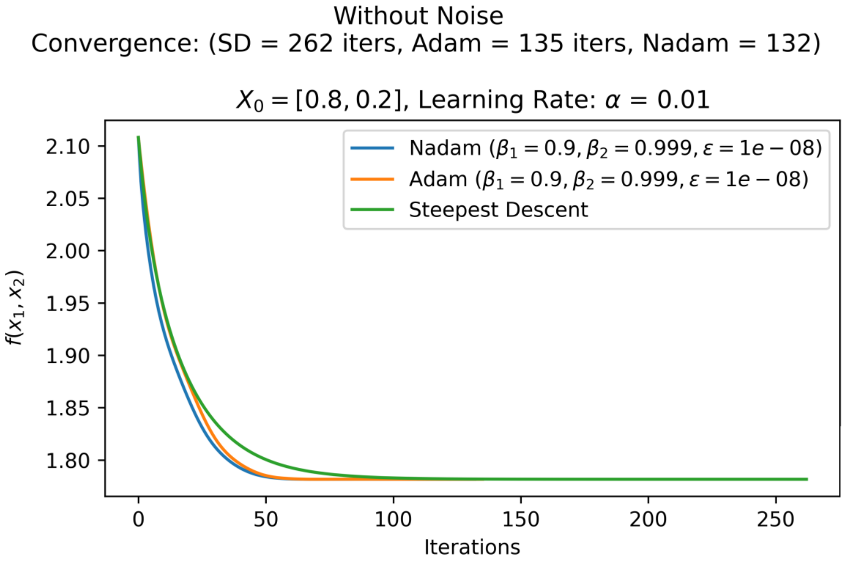




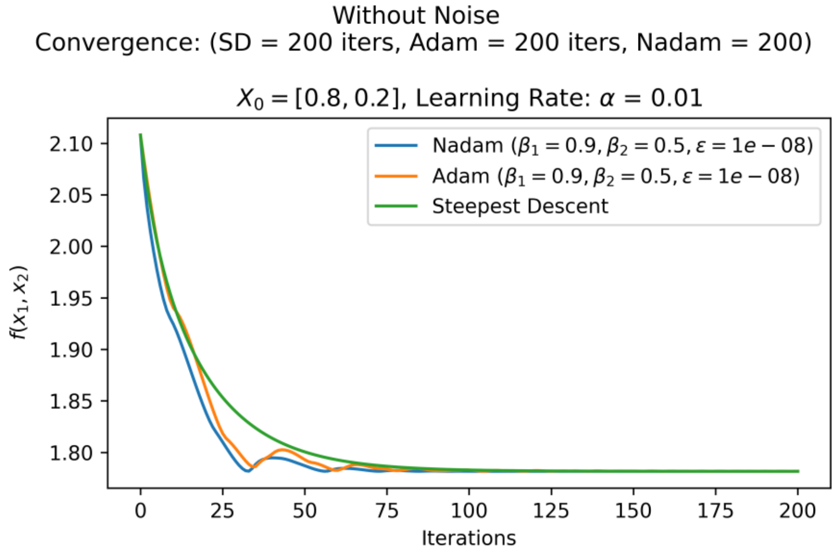
Effect of initial learning rate α

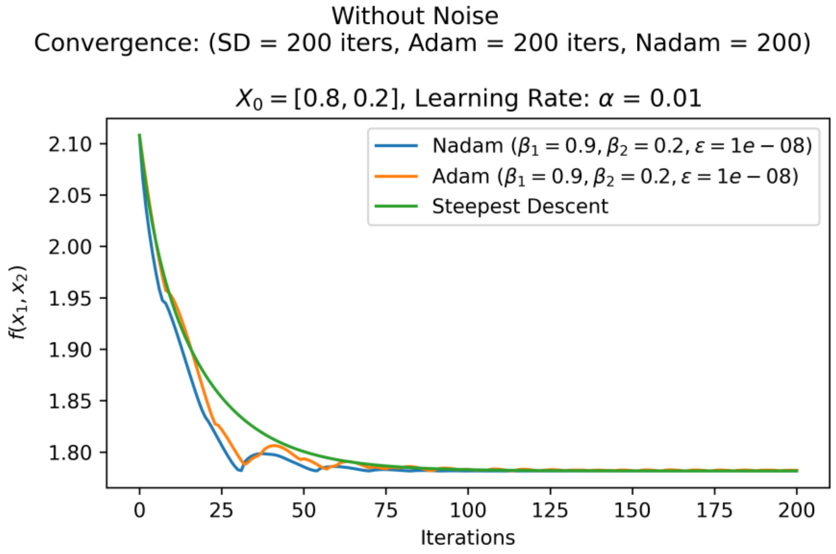
Numerical Results

Convergence Criterion: $x_i - x_{i-1} = 1e-5$,
& 200 iterations for $\beta_1, \beta_2 = (0.9, 0.5), (0.9, 0.2)$



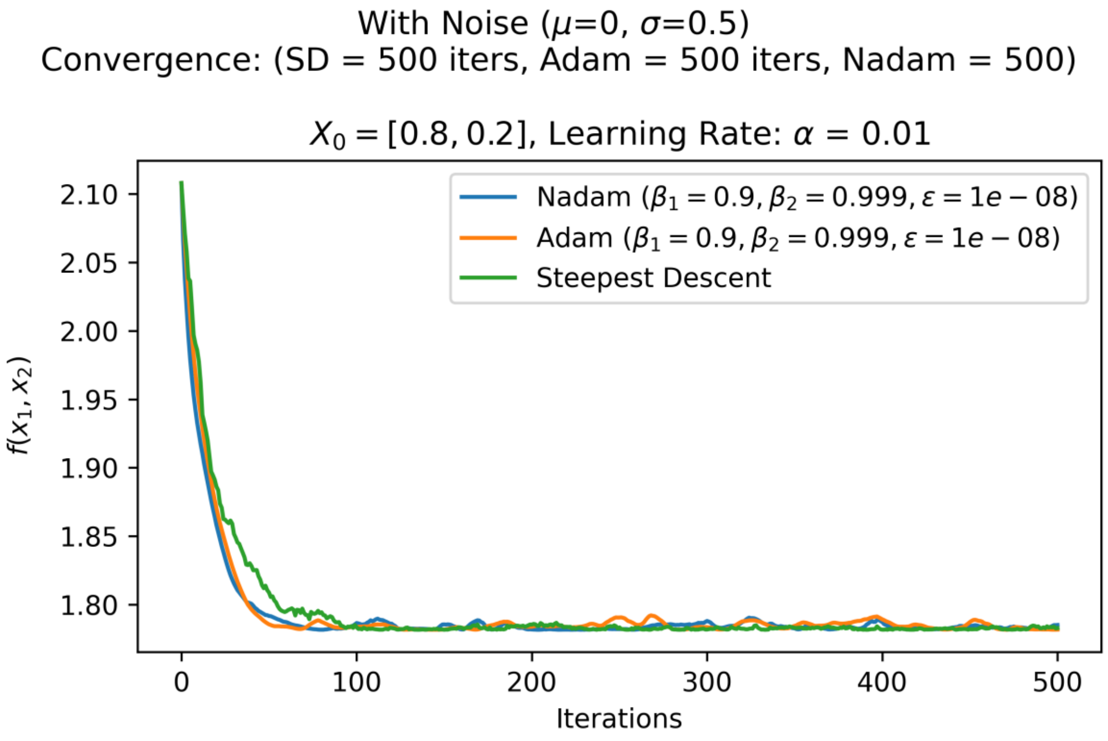
Hyperparameter Sensitivity

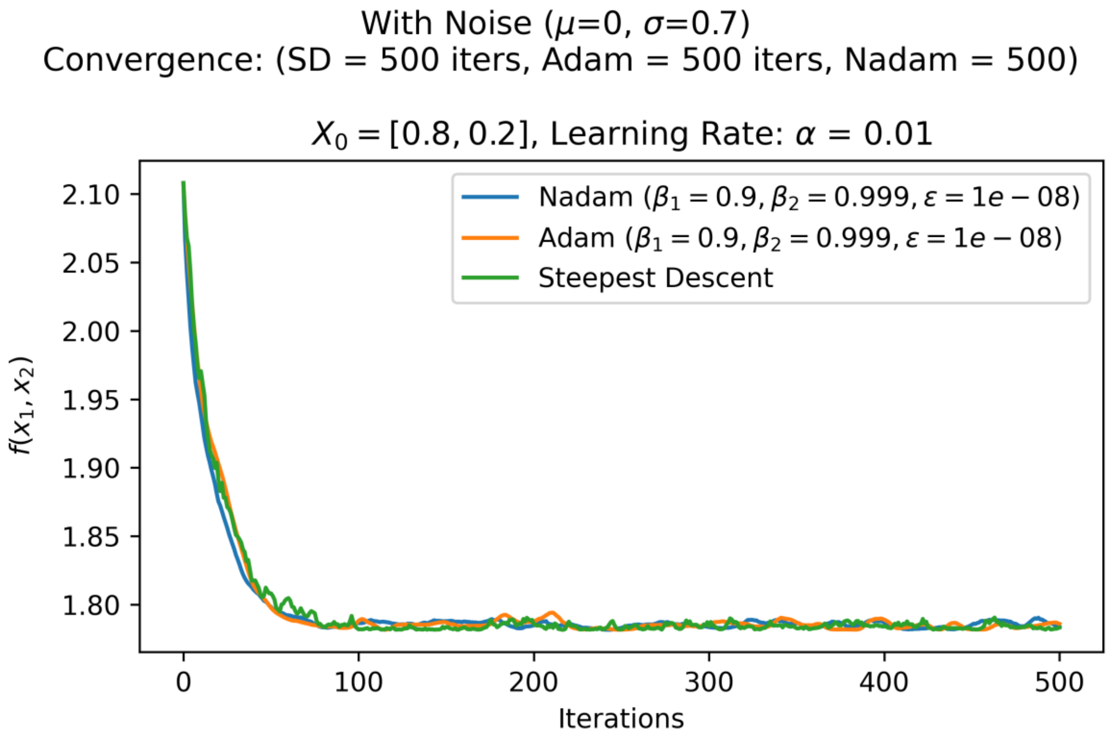




Numerical Results

Convergence Criterion: 500 iterations





Robustness to Noise

Conclusion

- $\underbrace{E[\nabla f(x)], E[\nabla (f(x)^2])}_{\substack{\textbf{Exponential Moving Average} \\ \textbf{(1st \& 2nd order moments)}}} \rightarrow \{\textbf{Convergence.}_{\text{Adam}}, \textbf{Convergence.}_{\text{Nadam}} >>>> \textbf{Convergence.}_{\text{SteepestDescent}}\}$

- $\textbf{Convergence}_{\text{Nadam}} > \textbf{Convergence}_{\text{Adam}}$:

  $\left.\boxed{\text{Step1: Momentum Cal.}} \rightarrow \boxed{\text{Step2: Gradient Calc.}}\right\}$ **High Quality Gradient**

- Both Nadam & Adam are **robust against noisy gradients**.
- Both Nadam & Adam are computationally expensive:
  - **Time Complexity:** Additional calculation of Expected values of the moments
  - **Space Complexity:** Maintaining previous values of moments which are *n-dimensional*
  - Nadam has a **slightly higher time complexity than Adam** due to additional gradient computation in the direction of the momentum step but is still **quicker to converge than Adam.**

References

1. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In Proc. 3rd International Conference on Learning Representations (ICLR) (ICLR, 2015).
2. Timothy Dozat. Incorporating Nesterov momentum into Adam. In International Conference on Learning Representations Workshops, 2016.

Q/A