Optimization Methods: Chart with Pros and Cons

Optimizer	Description	Pros	Cons
Gradient Descent (GD)	Updates parameters using the entire dataset at once.	- Guaranteed to move downhill (converges). - Simple to implement.	- Computationally expensive for large datasets. - May converge slowly.
Stochastic GD (SGD)	Updates parameters using one data point at a time.	- Computationally efficient for large datasets. - Can escape local minima.	- Highly noisy updates. - May lead to oscillations near the optimum.
Mini-batch GD	Updates parameters using a small subset (mini-batch) of data.	- Balances efficiency and noise reduction. - Faster convergence than SGD.	- Requires tuning of batch size. - Still computationally expensive for very large datasets.
Momentum (GD/SGD)	Adds a velocity term to smooth oscillations and accelerate convergence.	- Reduces oscillations in SGD. - Speeds up convergence.	- Requires tuning of the momentum parameter (β).

Optimizer	Description	Pros	Cons
RMSProp	Scales the learning rate by the square root of the gradient's EWMA.	- Handles sparse gradients efficiently. - Reduces oscillations significantly.	- May not converge to the global minimum. - Sensitive to hyperparameter tuning (β_2).
Adam (Adaptive Moment)	Combines Momentum (EWMA of gradients) and RMSProp (EWMA of squared gradients).	- Adaptive learning rate for each parameter. - Works well for sparse gradients.	- Sensitive to hyperparameters. - May not converge to the exact global minimum.
GD with Nesterov Momentum	Uses lookahead gradients to improve updates in momentum.	- Speeds up convergence further than vanilla momentum. - Improves smoothness in updates.	- More computationally intensive than simple momentum.

Comparison of Key Characteristics

Optimizer	Uses Entire Dataset?	Batch Size Used	Noise Level	Learning Rate Adaptive?	Gradient Averaging?
GD	Yes	Full	None	No	No
SGD	No	1 (single sample)	High	No	No
Mini-batch GD	No	User-defined	Moderate	No	No
GD with Momentum	Yes/Batch	Full/User-defined	Low	No	Momentum smoothing
RMSProp	No	User-defined	Moderate	Yes	Yes
Adam	No	User-defined	Moderate	Yes	Yes

Conclusion

Each optimizer has its strengths and weaknesses. Here's when to use each:

- **GD**: Small datasets, or when exact convergence is needed.
- **SGD**: Large datasets, where computational efficiency is critical.
- **Mini-batch GD**: A compromise between GD and SGD for medium to large datasets.

- **Momentum**: Use with GD or SGD to smooth and accelerate convergence.
- **RMSProp**: For non-convex problems or sparse gradients.
- **Adam**: General-purpose optimizer for most machine learning problems.