SGD with Momentum and Nesterov Acceleration

The Stochastic Gradient Descent (SGD) algorithm with Momentum and Nesterov Acceleration is a widely used optimization method in machine learning. Momentum helps accelerate convergence by accumulating a velocity vector in the direction of consistent gradients, while Nesterov acceleration improves stability and convergence by looking ahead in the direction of the velocity vector.

```
input : $\gamma$ (lr), $\theta_0$ (params), $f(\theta)$ (objective), $\lambda$ (weight decay),
        $\mu$ (momentum), $\tau$ (dampening), nesterov, maximize
-----------------------------------------------------------------
for $t = 1$ to $\ldots$ do
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
    if $\lambda \neq 0$
        $g_t \leftarrow g_t + \lambda\theta_{t-1}$
    if $\mu \neq 0$
        if $t > 1$
            $\mathbf{b}_t \leftarrow \mu\mathbf{b}_{t-1} + (1-\tau)g_t$
        else
            $\mathbf{b}_t \leftarrow g_t$
        if nesterov
            $g_t \leftarrow g_t + \mu\mathbf{b}_t$
        else
            $g_t \leftarrow \mathbf{b}_t$
    if maximize
        $\theta_t \leftarrow \theta_{t-1} + \gamma g_t$
    else
        $\theta_t \leftarrow \theta_{t-1} - \gamma g_t$
-----------------------------------------------------------------
return $\theta_t$
```

Explanation of Parameters and Variables

- γ (**lr**): The learning rate. It determines the step size of the parameter updates. A smaller learning rate leads to slower but more stable convergence.

- θ_0 (**params**): The initial parameters of the model. These are the values that the optimization algorithm will adjust to minimize the objective function.

- $f(\theta)$ (**objective**): The objective function (or loss function) that the algorithm aims to minimize. It is a function of the parameters θ.

- λ (**weight decay**): The weight decay coefficient. It adds a penalty proportional to the squared magnitude of the parameters to the objective function, encouraging smaller parameter values.

- μ (**momentum**): The momentum coefficient. It accelerates the optimization process by adding a fraction of the previous update to the current update. If $\mu = 0$, momentum is not used.

- τ (**dampening**): The dampening coefficient for momentum. It reduces the effect of momentum by scaling the current gradient before adding it to the velocity vector.

- **nesterov**: A boolean flag indicating whether to use Nesterov accelerated gradient (NAG). If enabled, the algorithm adjusts the gradient computation to look ahead in the direction of the momentum vector.

- **maximize**: A boolean flag indicating whether to maximize the objective function instead of minimizing it. If enabled, the algorithm updates parameters in the direction of the gradient (ascent) rather than against it (descent).
- g_t: The gradient of the objective function with respect to the parameters at time step t.
- $\mathbf{b}_t$: The momentum buffer (velocity vector) at time step t. It accumulates the gradients over time, scaled by the momentum coefficient μ.
- θ_t: The updated parameters at time step t.