Stochastic Gradient with Decreasing Step Sizes

- To get convergence, we need a decreasing step size.
  - Shrinks size of ball to zero so we converge to w^*.
- But it can't shrink too quickly:
  - Otherwise, we don't move fast enough to reach the ball.
- Stochastic gradient converges to a stationary point if:
  - Ratio of sum of squared step-sizes over sum of step-sizes converges to 0.

"how much noise affects you →

$$\frac{\sum_{t=1}^{\infty}(\alpha^t)^2}{\sum_{t=1}^{\infty}\alpha^t} = 0$$

"how far you can get" →

  - This choice also works for non-smooth funtions like SVMs.
    - Function must be continuous and not "too crazy" (we're still figuring it out for non-convex).

Stochastic Gradient with Decreasing Step Sizes

- For convergence step-sizes need to satisfy: $\sum_{t=1}^{\infty}(\alpha^t)^2 / \sum_{t=1}^{\infty}\alpha^t = 0$
- Classic solution is to use a step-size sequence like $\alpha^t = O(1/t)$.

$$\sum_{t=1}^{\infty}\alpha^t = \underbrace{\sum_{t=1}^{\infty}\frac{1}{t} = \infty}_{\text{"we can get everywhere"}} \qquad \sum_{t=1}^{\infty}(\alpha^t)^2 = \underbrace{\sum_{t=1}^{\infty}\frac{1}{t^2} < \infty}_{\text{"effect of variance goes to zero"}}$$

  - E.g., $\alpha^t = .001/t$.
- Unfortunately, this often works badly in practice:
  - Steps get really small really fast.
  - Some authors add extra parameters like $\alpha^t = \gamma/(\beta t + \Delta)$, which helps a bit.
  - One of the only cases where this works well: binary SVMs with $\alpha^t = 1/\lambda t$.

Stochastic Gradient with Decreasing Step Sizes

- How do we pick step-sizes satisfying $\sum_{t=1}^{\infty}(\alpha^t)^2 / \sum_{t=1}^{\infty}\alpha^t = 0$

- Better solution is to use a step-size sequence like $\alpha^t = O(1/\sqrt{t})$.

$$\sum_{t=1}^{K}\alpha^t = \sum_{t=1}^{K}\frac{1}{\sqrt{t}} = O(\sqrt{K}) \qquad \sum_{t=1}^{k}(\alpha^t)^2 = \sum_{t=1}^{k}\frac{1}{t} = O(\log k)$$

  - E.g., use $\alpha^t = .001/\sqrt{t}$
  - Both sequences diverge, but denominator diverges faster.

- This approach (roughly) optimizes rate that it goes to zero.
  - Better worst-case theoretical properties (and more robust to step-size).
  - Often better in practice too.

Stochastic Gradient with Constant Step Sizes?

- Alternately, could we just use a constant step-size.
  - E.g., use $\alpha^t = .001$ for all ‘t’.

- This will not converge to a stationary point in general.
  - However, do we need it to converge?

- What if you only care about the first 2-3 digits of the test error?
  - Who cares if you aren’t able to get 10 digits of optimization accuracy?

- There is a step-size small enough to achieve any fixed accuracy.
  - Just need radius of “ball” to be small enough.

A Practical Strategy for Deciding When to Stop

- In gradient descent, we can stop when gradient is close to zero.

- In stochastic gradient:
  - Individual gradients don’t necessarily go to zero.
  - We can’t see full gradient, so we don’t know when to stop.

- Practical trick:
  - Every ‘k’ iterations (for some large ‘k’), measure validation set error.
  - Stop if the validation set error “isn’t improving”.
    - We don’t check the gradient, since it takes a lot longer for the gradient to get small.
    - This “early stopping” can also reduce overfitting.