Review of Probability and Stochastic gradient descent (SGD)

October 2, 2018

1 Motivation

Given data $(x_i, y_i)_{i=1}^m$ ($x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$), we wanted to minimize

(1) (Linear-least squares) $\min_\theta \|X\theta - y\|_2^2 = \sum_{i=1}^m (x_i^T\theta - y_i)^2$

(2) (Logistic regression) $\min_\theta \sum_{i=1}^m \log(1 + \exp(\theta^T x_i)) - y_i\theta^T x_i$

Observe both of these take the form

$$\text{(Empirical loss/large finite sums)} \qquad \min_\theta f(\theta) = \frac{1}{m}\sum_{i=1}^m \ell_i(\theta) = \frac{1}{m}\sum_{i=1}^m \ell(\theta; (x_i, y_i)). \qquad (\star)$$

- Many ML problems take this form
- Think m is large- the number of samples is big
- Part of a more general form:

$$\text{(Expected loss)} \qquad \min_\theta f(\theta) = \mathbf{E}_\xi \ell(\theta; \xi) \qquad (\star\star)$$

  (i) ξ is a random variable

  (ii) $\ell(\theta; \xi)$ is the loss of a predictor θ on a sample ξ

  (iii) More interested in $(\star\star)$ but can't solve it so we approx. it by empirical loss $(\star)$

Problem: Because m is large, we can't compute $\nabla f(\theta)$ or even evaluate $f(\theta)$.

Remark 1. For $(\star\star)$, convexity/differentiability (resp.) means $\ell(\cdot; \xi)$ is convex/differentiable for all ξ

Question: How do we minimize $(\star)$ and $(\star\star)$?

$\Rightarrow$ Use probability! Instead of computing $\nabla f(\theta)$ we randomly choose sample x_i (or ξ_i) and use $\nabla \ell_i(\theta)$ or $\nabla \ell(\theta; \xi_i)$ as a surrogate for the full gradient.

2 Review of Probability

2.1 Random variables, expectation, and variance

We will denote Ω the underlying space, the collection of sets $\mathcal{F}$ a *σ-algebra* on Ω, and the probability measure as $\mathbf{Pr} : \mathcal{F} \to [0, 1]$. A *probability space* is $(\Omega, \mathcal{F}, \mathbf{Pr})$. The *indicator of a set* A, $1_A : \Omega \to \mathbb{R}$ is

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \in A^c \end{cases}$$

Definition 2.1 (Random variable)**.** A function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B})$ is a *random variable/vector* if X is a measurable map. We say the σ-algebra generated by the random variable X, denoted as $\sigma(X)$, to be the smallest σ-algebra for which X is measurable. In particular, this means

$$\sigma(X) = \{\{X \in B\} \, : \, B \in \mathcal{B}\}.$$

The *expected value* of a random variable $X : (\Omega, \mathcal{F}, \mathbf{Pr}) \to (\mathbb{R}^n, \mathcal{B})$ is $\mathbf{E}[X] = \int_\Omega X(\omega) \, d\mathbf{Pr}$. Properties of integrals such convergence results, linearity, etc hold for the expected value. The *variance* of a random variable is

$$\mathbf{E}[(X - \mathbf{E}[X])^2] = \mathrm{Var}(X).$$

An important inequality often used to related variance to the probability of an event occurring is *Chebyshev's or Markov inequality*

$$a^2 \mathbf{Pr}(|X| \geq a) \leq \mathbf{E}X^2.$$

2.2 Different types of convergence for probabilistic algorithms

There are many different types of convergence results one encounters. I will list a subset of them for which we often see in optimization:

- *"with high probability, w.h.p"*: An event A occurs with high probability with respect to parameter n if it occurs with probability p_n and $\lim_{n\to\infty} p_n = 1$.
- *"almost surely (a.s) or almost everywhere"*: Suppose $\{X_n\}$ is a sequence of random variables (*e.g.* generated by a stochastic algorithm) then we say $X_n \to X$ a.s if $\mathbf{Pr}(X_n \to X) = 1$. In optimization, we might be interested if an algorithm produces iterates such that $\nabla f(X_n) \to 0$ a.e.

2.3 Conditional Expectation

Definition 2.2 (Conditional Expectation)**.** Given a probability space $(\Omega, \mathcal{F}_0, \mathbf{Pr})$, a σ-algebra $\mathcal{F} \subset \mathcal{F}_0$, and a random variable $X \in \mathcal{F}_0$ with $\mathbf{E}[|X|] < \infty$. We define the *conditional expectation of* X given a σ-algebra $\mathcal{F}$, $\mathbf{E}[X|\mathcal{F}]$ to be any random variable Y that has

(i) $Y \in \mathcal{F}$, i.e. Y is $\mathcal{F}$ measurable

(ii) for all $A \in \mathcal{F}$, $\int_A X\, dP = \int_A Y\, dP$

Alternatively, one can define $\mathbf{E}[X|\mathcal{F}]$ (under the additional assumption that $\mathbf{E}[X^2] < \infty$) as

$$\mathbf{E}[X|\mathcal{F}] = \text{argmin}_{Y \in L^2(\mathcal{F})}\mathbf{E}[(X-Y)^2].$$

It is important to note that the conditional expectation is a random variable itself. Intuitively, we think of the σ-algebra $\mathcal{F}$ as describing the information we have at our disposal- for each $A \in \mathcal{F}$, we know whether or not A has occurred. Thus, $\mathbf{E}[X|\mathcal{F}]$ is then our "best guess" of the value of X given the information we have.

- Example (complete knowledge) If $X \in \mathcal{F}$, then $\mathbf{E}[X|\mathcal{F}] = X$ i.e. if we know X then our best guess of X is X itself. In particular, if $X = c$, a constant, then $\mathbf{E}[c|\mathcal{F}] = c$.

- Example (no information) Suppose X is independent of $\mathcal{F}$ i.e for all $B \in \mathcal{B}$ and $A \in \mathcal{F}$

  $$\mathbf{Pr}(\{X \in B\} \cap A) = \mathbf{Pr}(\{X \in B\})\mathbf{Pr}(A).$$

  We claim, in this case, $\mathbf{E}[X|\mathcal{F}] = \mathbf{E}[X]$ (i.e. if you don't know anything about X, then the best guess is the mean $\mathbf{E}[X]$).

- Example (undergraduate definition). In undergraduate, we were taught for any two set A, B such that $\mathbf{Pr}(B) \neq 0$,

  $$\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}.$$

  Suppose $\Omega_1, \Omega_2, \ldots$ is a finite or infinite partition of Ω into disjoint sets, each with positive probability and let $\mathcal{F} = \sigma(\Omega_1, \Omega_2, \ldots)$. Then

  $$\mathbf{E}[X|\mathcal{F}] = \sum_{i=1}^{\infty} 1_{\Omega_i} \frac{\mathbf{E}[X; \Omega_i]}{\mathbf{Pr}(\Omega_i)}$$

  If A and B ($\mathbf{Pr}(B) > 0$) are sets and $X = 1_A$, then

  $$\mathbf{Pr}(A|\sigma(B)) = \mathbf{E}[1_A|\sigma(B) = \{B, B^c, \Omega, \varnothing\}] = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)} 1_B + \frac{\mathbf{Pr}(A \cap B^c)}{\mathbf{Pr}(B^c)} 1_{B^c}.$$

  Remark 2. Let X, Y be two random variables. We often use the notation

  $$\mathbf{E}[X|Y] = \mathbf{E}[X|\sigma(Y)].$$

- Suppose X and Y are independent. Let φ be a function with $\mathbf{E}[\varphi(X,Y)] < \infty$ and let $g(x) = \mathbf{E}[\varphi(x,Y)]$. Then

  $$\mathbf{E}[\varphi(X,Y)|X] = g(X).$$

We list some important properties of the conditional expectation:

- (Linearity) $\mathbf{E}[aX + Y|\mathcal{F}] = a\mathbf{E}[X|\mathcal{F}] + \mathbf{E}[Y|\mathcal{F}]$

- If $X \leq Y$, then $\mathbf{E}[X|\mathcal{F}] \leq \mathbf{E}[Y|\mathcal{F}]$
- If $X_n \geq 0$ and $X_n \uparrow X$ with $\mathbf{E}[X] < \infty$, then $\mathbf{E}[X_n|\mathcal{F}] \uparrow \mathbf{E}[X|\mathcal{F}]$.
- (Jensen's inequality) If φ is convex and $\mathbf{E}|X| < \infty$ and $\mathbf{E}|\varphi(X)| < \infty$, then $\varphi(\mathbf{E}[X|\mathcal{F}]) \leq \mathbf{E}[\varphi(X)|\mathcal{F}]$.
- $\mathbf{E}[\mathbf{E}[Y|\mathcal{F}]] = \mathbf{E}[Y]$.
- If $\mathcal{F} \subset \mathcal{G}$ and $\mathbf{E}[X|\mathcal{G}] \in \mathcal{F}$ then $\mathbf{E}[X|\mathcal{F}] = \mathbf{E}[X|\mathcal{G}]$.
- (Tower rule) If $\mathcal{F}_1 \subset \mathcal{F}_2$, then $\mathbf{E}[\mathbf{E}[X|\mathcal{F}_1]|\mathcal{F}_2] = \mathbf{E}[X|\mathcal{F}_1]$ and $\mathbf{E}[\mathbf{E}[X|\mathcal{F}_2]|\mathcal{F}_1] = \mathbf{E}[X|\mathcal{F}_1]$.
- If $X \in \mathcal{F}$ and $\mathbf{E}|Y| < \infty$ and $\mathbf{E}|XY| < \infty$, then

$$\mathbf{E}[XY|\mathcal{F}] = X\mathbf{E}[Y|\mathcal{F}].$$

3 Stochastic Gradient Descent (SGD)

This is a very old method established by Robbins and Monro [1951]: gradients do not need to be computed exactly in order to guarantee progress toward optimum. They observed that as long as the gradients are correct *on average*, the error introduced by the gradient approximations will eventually vanish.

Recall we are interested in solving

$$\min_x f(x) := \mathbf{E}_\xi[\ell(x;\xi)] \quad \text{and} \quad \min_x \ f(x) := \frac{1}{m}\sum_{i=1}^m f_i(x) = \frac{1}{m}\sum_{i=1}^m \ell(x,\xi_i).$$

For simplicity, lets assume $f(x)$ is has L- smooth.

What optimization algorithm might we run?-gradient descent

$$x_{k+1} = x_k - \alpha_k \tilde{g}(x_k,\xi_k), \quad \text{where } \tilde{g}(x_k,\xi_k) \text{ is an estimator of the gradient.}$$

Definition 3.1 (Stochastic oracle)**.** A *stochastic oracle* for a convex function $f : \mathbb{R}^m \to \mathbb{R}^n$ inputs a random variable X and outputs a random variable $\tilde{g}(X)$ such that $\mathbf{E}[\tilde{g}(X)|\sigma(X)] \in \partial f(X)$ ($\mathbf{E}[\tilde{g}(X)|\sigma(X)] = \nabla f(X)$). This is an *unbiased estimator* of gradient.

Question: How to do we select $\tilde{g}(X)$?

Under natural technical conditions,
First, let's check that in both cases we are getting an unbiased estimator of the gradient:

- (Empirical risk) $\mathbf{E}[\nabla f_I(X_k)|\sigma(X_k)] = \frac{1}{m}\sum_{i=1}^m \nabla f_i(X_k)$.

Algorithm 1: Stochastic gradient descent for expected loss

initialize: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in \mathbb{R}$ (not random)
for $k = 0, 1, \ldots$
Draw ξ randomly from the unknown distribution
Compute $\nabla_x \ell(x_k, \xi)$
Set $x_{k+1} = x_k - \alpha_k \nabla_x \ell(x_k, \xi)$
Update α_{k+1}
end

Algorithm 2: Stochastic gradient descent for empirical risk

initialize: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in \mathbb{R}$ (not random)
for $k = 0, 1, \ldots$
Draw uniformly at random $I \in [m]$
Compute $\nabla f_I(x)$
Set $x_{k+1} = x_k - \alpha_k \nabla f_I(x)$
Update α_{k+1}
end

- (Expected loss) Under some technical assumptions (ξ and X are independent and interchanging of derivatives), $\mathbf{E}[\nabla \ell(X_k; \xi) | \sigma(X_k)] = \nabla \mathbf{E}_\xi[\ell(\cdot; \xi)](X_k)$

What can go wrong?

- Stepsize- fixed or decreasing. We can not expect a fixed stepsize because even if at the optimal we will still take a step.
- Control the fluctuations of $\tilde{g}(x)$- (nonsmooth case) we will have to assume $\mathbf{E}[\|\tilde{g}(X)\|^2] < B^2$ and in the smooth case $\mathbf{E}[\|\tilde{g}(X) - \nabla f(X)\|^2] \leq \sigma^2$
- Difference between Algorithm 1 and Algorithm 2: Suppose we only have m iid samples $\xi_1, \ldots, \xi_m$ and we run Algorithm 1. How many times can we run Algorithm 1? How many times can we run Algorithm 2? For Algorithm 1, we can only run it m times. Why? Consider X_{m+1}. Then $\sigma(X_{m+1})$ contains information about $\xi_1, \ldots, \xi_m$. Therefore, if I generate my next ξ it will not be independent of X_{m+1} and thus the conditional expectation doesn't match. Whereas in Algorithm 2 can run it infinitely often.

Class	Stepsize	Rate
L-smooth, $\mathbf{E}\|\nabla f(X) - \tilde{g}(X)\|^2 \leq \sigma^2$	$\min\{1/L, \frac{\varepsilon^2}{L\sigma}\}$	$\mathbf{E}[\|\nabla f(X_k)\|^2] \leq \varepsilon^2$
convex (not diff), $\mathbf{E}[\|\tilde{g}(X)\|^2] \leq B^2$	$\frac{R}{B}\sqrt{\frac{2}{T}}$	$\mathbf{E}\left[f\left(\frac{1}{T}\sum_{k=1}^T X_k\right)\right] - f^* \leq RB\sqrt{\frac{2}{T}}$
ℓ-strongly cvx (not diff), $\mathbf{E}[\|\tilde{g}(X)\|^2] \leq B^2$	$\frac{2}{\ell(k+1)}$	$\mathbf{E}\left[f\left(\sum_{k=1}^T \frac{2k}{T(T+1)} x_k\right)\right] - f^* \leq \frac{2B^2}{\ell(k+1)}$
cvx, L smooth, $\mathbf{E}\|\nabla f(X) - \tilde{g}(X)\|^2 \leq \sigma^2$	$\frac{1}{L+1/\eta}$, $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{T}}$	$\mathbf{E}f\left(\frac{1}{T}\sum_{k=1}^T x_k\right) - f^* \leq R\sigma\sqrt{\frac{2}{T}} + \frac{LR^2}{T}$
ℓ strong cvx, L-smooth, $\mathbf{E}\|\nabla f(X) - \tilde{g}(X)\|^2 \leq \sigma^2$	$\frac{2}{\ell\sigma(1+k)}$	$\mathbf{E}[f(X_k) - f^*] \leq O(1)\frac{1}{1+k}$

Here $R = \|x_1 - x^*\|$.

Theorem 3.2 (Convex, SGD convergence)**.** Suppose f is convex and attains minimum at x^*. (You don't necessarily need f to attain its minimum, but then you need to have a bounded domain.) Assume $\mathbf{E}[\|\tilde{g}(X)\|^2] \leq B^2$ [bound on the fluctuations]. Then after T iterations of *SGD* with $\alpha = \frac{R}{B}\sqrt{\frac{2}{T}}$ satisfies

$$\mathbf{E}\left[f\left(\frac{1}{T}\sum_{k=1}^{T} X_k\right)\right] - f^* \leq RB\sqrt{\frac{2}{T}}$$

where $R = \|x_1 - x^*\|$.

Remark 3. In fact, the proof of this theorem shows that for a fixed step-size, α, SGD converges to a ball. Moreover, the same argument with the stepsize α_k varying on each iteration shows that one should take $\alpha_k \approx 1/k$ in order to guarantee convergence.

Proof. By convexity, we have

$$\begin{aligned}
f(X_k) - f^* &\leq \nabla f(X_k)^T (X_k - x^*) \\
\text{add/sub. } \tilde{g}(X_k) \quad &= \tilde{g}(X_k)^T (X_k - x^*) + (\nabla f(X_k) - \tilde{g}(X_k))^T (X_k - x^*) \\
\text{defn of } \tilde{g}(X_k) = 1/\alpha(X_k - X_{k+1}) \quad &= \frac{(X_k - X_{k+1})^T (X_k - x^*)}{\alpha} + \tilde{g}(X_k)^T (X_k - x^*) \\
&\quad + (\nabla f(X_k) - \tilde{g}(X_k))^T (X_k - x^*) \\
2(a-b)^T(a-c) = \|a-b\|^2 + \|a-c\|^2 - \|b-c\|^2 \quad &= \frac{1}{2\alpha}\left(\|X_k - x^*\|^2 - \|X_{k+1} - x^*\|^2\right) + \frac{1}{2\alpha}\|X_k - X_{k+1}\|^2 \\
&\quad + (\nabla f(X_k) - \tilde{g}(X_k))^T (X_k - x^*) \\
\text{plug in } X_{k+1} = X_k - \alpha\tilde{g}(X_k) \quad &= \frac{1}{2\alpha}\left(\|X_k - x^*\|^2 - \|X_{k+1} - x^*\|^2\right) + \frac{\alpha}{2}\|\tilde{g}(X_k)\|^2 \\
&\quad + (\nabla f(X_k) - \tilde{g}(X_k))^T (X_k - x^*).
\end{aligned}$$

Next we take conditional expectation w.r.t. to $\sigma(X_k)$ and obtain

$$\begin{aligned}
\mathbf{E}[f(X_k) - f^* | \sigma(X_k)] \leq \frac{1}{2\alpha}\left(\mathbf{E}[\|X_k - x^*\|^2 | \sigma(X_k)] - \mathbf{E}[\|X_{k+1} - x^*\|^2 | \sigma(X_k)]\right) + \frac{\alpha}{2}\mathbf{E}[\|\tilde{g}(X_k)\|^2 | \sigma(X_k)] \\
+ \mathbf{E}[(\nabla f(X_k) - \tilde{g}(X_k))^T (X_k - x^*) | \sigma(X_k)].
\end{aligned}$$

We note that

$$\begin{aligned}
\mathbf{E}[(\nabla f(X_k) - \tilde{g}(X_k))^T (X_k - x^*) | \sigma(X_k)] &= \mathbf{E}[\nabla f(X_k) - \tilde{g}(X_k) | \sigma(X_k)]^T (X_k - x^*) \\
&= 0.
\end{aligned}$$

Take expectations and use total law of expectation,

$$\mathbf{E}[f(X_k) - f^*] \leq \frac{1}{2\alpha}\left(\mathbf{E}[\|X_k - x^*\|^2] - \mathbf{E}[\|X_{k+1} - x^*\|^2]\right) + \frac{\alpha B^2}{2}.$$

Summing up we get

$$\frac{1}{T}\sum_{k=1}^{T} \mathbf{E}[f(X_k)] - f^* \leq \frac{1}{2\alpha T}\|x_1 - x^*\|^2 + \frac{\alpha}{2}B^2,$$

minimizing the RHS with respect to α gives the result. □

Theorem 3.3 (cvx, L-smooth)**.** *Assume that* $\mathbf{E}[\|\nabla f(X) - \tilde{g}(X)\|^2] \leq \sigma^2$*. Then with step-size* $\frac{1}{L+1/\eta}$ *and* $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{T}}$ *satisfies*

$$\mathbf{E}[f\left(\frac{1}{T}\sum_{k=1}^{T} x_{k+1}\right)] - f^* \leq R\sigma\sqrt{\frac{2}{T}} + \frac{LR^2}{T},$$

where $R = \|x_1 - x^*\|$.

Proof. By L-smoothness, we have

$$\begin{aligned}
f(X_{k+1}) - f(X_k) &\leq \nabla f(X_k)^T(X_{k+1} - X_k) + \frac{L}{2}\|X_{k+1} - X_k\|^2 \\
\text{add/sub. } \tilde{g}(X_k) \quad &= \tilde{g}(X_k)^T(X_{k+1} - X_k) + (\nabla f(X_k) - \tilde{g}(X_k))^T(X_{k+1} - X_k) + \frac{L}{2}\|X_{k+1} - X_k\|^2 \\
\text{complete the square} \quad &\leq \tilde{g}(X_k)^T(X_{k+1} - X_k) + \frac{\eta}{2}\|\nabla f(X_k) - \tilde{g}(X_k)\|^2 + \frac{1}{2}(L + \frac{1}{\eta})\|X_{k+1} - X_k\|^2 \\
\text{add/sub. } x^* \quad &\leq \tilde{g}(X_k)^T(X_{k+1} - x^*) + \tilde{g}(X_k)^T(x^* - X_k) \\
&\quad + \frac{\eta}{2}\|\nabla f(X_k) - \tilde{g}(X_k)\|^2 + \frac{1}{2}(L + \frac{1}{\eta})\|X_{k+1} - X_k\|^2 \\
\text{as before } (a-b)^T(a-c) \quad &= \frac{1}{2\alpha}\left(\|X_k - x^*\|^2 - \|X_{k+1} - x^*\|^2\right) - \frac{1}{2\alpha}\|X_{k+1} - X_k\|^2 + \tilde{g}(X_k)^T(x^* - X_k) \\
&\quad + \frac{\eta}{2}\|\nabla f(X_k) - \tilde{g}(X_k)\|^2 + \frac{1}{2}(L + \frac{1}{\eta})\|X_{k+1} - X_k\|^2.
\end{aligned}$$

Choose α to cancel out the term $\|X_{k+1} - X_k\|^2$ (i.e. $\alpha = 1/(L + 1/\eta)$). Hence, we have that

$$\begin{aligned}
f(X_{k+1}) &\leq f(X_k) + \tilde{g}(X_k)^T(x^* - X_k) + \frac{1}{2\alpha}\left(\|X_k - x^*\|^2 - \|X_{k+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(X_k) - \tilde{g}(X_k)\|^2 \\
\text{cvx} \quad &\leq f^* + (\tilde{g}(X_k) - \nabla f(X_k))^T(x^* - X_k) + \frac{1}{2\alpha}\left(\|X_k - x^*\|^2 - \|X_{k+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(X_k) - \tilde{g}(X_k)\|^2
\end{aligned}$$

Next we take conditional expectation w.r.t. to $\sigma(X_k)$ and obtain

$$\begin{aligned}
\mathbf{E}[f(X_{k+1}) - f^*|\sigma(X_k)] \leq \frac{1}{2\alpha}\Big(&\mathbf{E}[\|X_k - x^*\|^2|\sigma(X_k)] - \mathbf{E}[\|X_{k+1} - x^*\|^2|\sigma(X_k)]\Big) + \frac{\eta}{2}\mathbf{E}[\|\nabla f(X_k) - \tilde{g}(X_k)\|^2|\sigma(X_k)] \\
&+ \mathbf{E}[(\nabla f(X_k) - \tilde{g}(X_k))^T(x^* - X_k)|\sigma(X_k)].
\end{aligned}$$

We note that

$$\mathbf{E}[(\nabla f(X_k) - \tilde{g}(X_k))^T(x^* - X_k)|\sigma(X_k)] = \mathbf{E}[\nabla f(X_k) - \tilde{g}(X_k)|\sigma(X_k)]^T(x^* - X_k) = 0.$$

Using the total law of expectation and the variance bound, we obtain

$$\mathbf{E}[f(X_{k+1}) - f^*] \leq \frac{1}{2\alpha}\left(\mathbf{E}[\|X_k - x^*\|^2] - \mathbf{E}[\|X_{k+1} - x^*\|^2]\right) + \frac{\eta\sigma^2}{2}.$$

Summing up we get

$$\frac{1}{T}\sum_{k=1}^{T}\mathbf{E}[f(X_{k+1})] - f^* \leq \frac{1}{2\alpha T}\|x_1 - x^*\|^2 + \frac{\eta\sigma^2}{2} = \frac{L + 1/\eta}{2T}\|x_1 - x^*\|^2 + \frac{\eta\sigma^2}{2},$$

minimizing the RHS with respect to η gives the result. □

Question: The convergence rate for SGD is slow compared to full gradient descent, what could we do to improve this rate?

We could add more samples. What does this do...reduces the variance. Suppose instead of uniformly choosing index $I \in [m]$, we choose $|S|$ number of indices in $[m]$:

$$x_{k+1} = x_k - \alpha_k \frac{1}{|S|} \sum_{i \in S} \tilde{g}_i(X_k).$$

When we do this, we reduce the variance σ^2, particularly,

$$\begin{aligned}\mathbf{E}[\|\frac{1}{|S|}\sum_{i\in S}\tilde{g}_i(X) - \nabla f(X)\|^2] &= \frac{1}{|S|^2}\mathbf{E}[\sum_{i\in S}(\tilde{g}_i(X) - \nabla f(X))\|^2] \\ \text{independence} \quad &= \frac{1}{|S|}\mathbf{E}[\|\tilde{g}_1(X) - \nabla f(X)\|^2] \leq \frac{2B^2}{|S|} =: \sigma^2_{\text{mini}}\end{aligned}$$

Great! We reduced variance, but do we lose anything? Yes, we do. Each iteration is much more expense. Suppose we can only call the stochastic oracle T times. Thus, it is $T/|S|$ iterations of minibatch we can call. Setting $T/|S|$ for T and σ_{mini} into the our rate, we obtain

$$R\sqrt{\frac{2B^2}{|S|}}\sqrt{\frac{2}{T/|S|}} + \frac{LR^2}{T/|S|} = 2\frac{RB}{\sqrt{T}} + \frac{|S|LR^2}{T},$$

so it doesn't pay in this case to use minibatching. The speed up you get the optimization problem due to reducing variance does not outweigh the cost.

3.1 Practical consideration

- minibatching may be beneficial if the cost per iteration to compute the minibatch is reduced for e.g. distribute across multiple processors. But, your minibatch $|S| << m$.
- Although SGD converges with stepsize decreasing, people generally run SGD until they see it stabilize. Then they shrink the step size parameter by some constant.

3.2 Why does SGD work so well in ML? Possible explanations

Intuitive Motivations:

- SGD uses information more efficiently! Consider a training set, called $\mathcal{S}$, which consists of ten copies of the set $\mathcal{S}_0$. A minimizer of the empirical risk over the larger set $\mathcal{S}$ is clearly a minimizer of the smaller set $\mathcal{S}_0$. If we apply "batch" approach to minimize the empirical risk, each iteration would be *10* times more expensive. SGD performs the same computation in both scenarios.
- Convergence rate in strongly convex case. SGD has a sublinear convergence of $1/\varepsilon$ where as full gradient descent has convergence $\log(1/\varepsilon)$ but costs m number of stochastic gradients. The total number cost of full gradient is $m\log(1/\varepsilon)$. However if m is very large then these might be comparable.
- Observed that SGD "generalizes" well. Namely SGD converges to "better optimum" then other more advance methods.