1 Setup

Previously we have discussed gradient methods for convex optimization in the deterministic setting, where everything may be calculated precisely. However, we would like to apply function optimization techniques to cases where the gradient oracle may not be exact.

In parametric machine learning, we would like to minimize a loss function over some parameter space. However, the data is random, so we want to minimize the expected loss over the data. Since many useful loss functions are convex in machine learning, we would like to build a theory for optimizing convex functions given a *stochastic oracle* for the gradient of the loss function. Specifying a stochastic oracle over the data is one way to add uncertainty into a model.

In this section we will focus on unbiased oracles for the gradient: That is, the expected value of the gradient is in the subdifferential set for the function: $\mathbb{E}\left[\tilde{g}(x)\right] \in \partial f(x)$. If the query points x are also random variables, we condition on x first: $\mathbb{E}\left[\tilde{g}(x)|x\right] \in \partial f(x)$. That is to say, the randomness over which the expectation is defined is the stochasticity of the gradient oracle, which is due to the fact that a gradient of the loss function is only sampled and not fully calculated (the reason for sampling may be due to the fact that we want to save time and not calculate everything, or may be due to the fact that the data is noisy and thus the gradient is noisy, etc.).

There are two simple oracle models which we can follow:

(a) **Non-separable stochastic oracle.** In this case, we define a stochastic oracle over the whole dataset all at once. We say

$$f(x) = \mathbb{E}_{\zeta}\left[l(x, \zeta)\right]$$

where x are the parameters of the model we are trying to learn and ζ is a specific datapoint. We assume that the loss function l is differentiable and convex. Our stochastic oracle can draw ζ from the unknown data distribution and report $\nabla_x l(x, \zeta)$.

(b) **Separable stochastic oracle.** The difficulty with the first definition is that the expectation is defined over the whole dataset all at once. That means if we were to expand the expectation as a sum, every data point would contribute exactly once to the value of $f(x)$. This reliance on defining $f(x)$ in terms of the expected loss means that the oracle is only unbiased for a single pass over the dataset. We can remedy this by defining $f(x)$ in a separable way:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

where $f_i(x) = l(x, \zeta_i)$. Here, the stochastic oracle may be defined by uniformly selecting some index $j \in [m]$ and reporting $\nabla f_j(x)$. Here, we have defined the loss *in terms of* the dataset (which is why it is called the *empirical loss*). This definition allows us to take as many passes over the dataset as we would like without affecting the bias of the gradient estimator. This setup is important for the online learning framework.

However, the requirement that the gradient is unbiased is not sufficient for obtaining convergence rates for optimization. We also need to make assumptions about the variance of the gradient estimator. We split the cases into smooth and unsmooth:

(a) Unsmooth: $\mathbb{E}\left[\|\tilde{g}(x)\|_*^2\right] \leq B^2$ for all x

(b) Smooth: $\mathbb{E}\left[\|\tilde{g}(x) - \nabla f(x)\|_*^2\right] \leq \sigma^2$ for all x

2 Non-smooth case

The essential punchline for the non-smooth case is that the results from nonstochastic convex optimization carry over due to the fact that our methods for non-smooth optimization use small step sizes. Recall

Definition 2.1. Stochastic Mirror Descent.
Begin with $x_1 \in \text{argmin}_{\mathcal{X}\cap\mathcal{D}}\phi(x)$. Recall that *mirror map* $\phi : \mathcal{D} \to \mathbb{R}$ is strictly convex and differentiable, $\nabla\phi$ is onto $\mathbb{R}^n$, and $\|\nabla\phi\|$ diverges on the boundary of $\mathcal{D}$. Then make the following update:

$$x_{t+1} = \text{argmin}_{x\in\mathcal{X}\cap\mathcal{D}}\eta\tilde{g}(x_t)^T x + D_\phi(x, x_t)$$

Recall that D_ϕ is the Bregman divergence $D_f(x, y) = f(x) - f(y) - \nabla f(y)^T(x - y)$, which satisfies $(\nabla f(x) - \nabla f(y))^T(x - z) = D_f(x, y) + D_f(z, x) - D_f(z, y)$.

Alternatively, we can describe stochastic mirror descent as following the following updates:

$$\nabla\phi(y_{s+1}) = \nabla\phi(x_s) - \eta\tilde{g}_s$$

$$x_{s+1} \in \Pi_{\mathcal{X}}^{\phi}(y_{s+1})$$

The only modification we make to standard mirror descent is replacing the gradient g_s with a gradient estimator $\tilde{g}_s$. Recall that we can interpret this step as attempting to minimize local linearization of the function while not stepping too far away from the previous point, with distance measured in terms of the Bregman divergence of the mirror map.

We now proceed to find convergence rates for this estimator in the non-smooth case.

Theorem 2.2. *Let ϕ be a mirror map 1-strongly convex on $\mathcal{X}\cap\mathcal{D}$ with respect to $\|\cdot\|$, and let $R^2 = \sup_{x\in\mathcal{X}\cap\mathcal{D}}\phi(x) - \phi(x_1)$. Let f be convex. Furthermore assume that $\mathbb{E}\left[\|\tilde{g}(x)\|_*^2 \leq B^2\right]$. Then stochastic mirror descent with $\eta = \frac{R}{B}\sqrt{2}t$ satisfies*

$$\mathbb{E}\left[f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right)\right] - \min_{x\in\mathcal{X}} f(x) \leq RB\sqrt{\frac{2}{t}}$$

Proof. We follow the proof of mirror descent's convergence with a few modifications. Since this was covered in Ch. 4, we sketch the beginning.

We begin by bounding the desired quantity in terms of $\tilde{g}_s^T(x_s - x)$. We will then be able to bound the sum over s of this quantity. We have

$$\begin{aligned}
\mathbb{E}\left[f(\frac{1}{t}\sum_{s=1}^{t} x_s)\right] - f(x) &\leq \frac{1}{t}\mathbb{E}\left[\sum_{s=1}^{t}(f(x_s) - f(x))\right] \text{ by convexity of } f \\
&\leq \frac{1}{t}\mathbb{E}\left[\sum_{s=1}^{t}\mathbb{E}\left[\tilde{g}(x_s)|x_s\right]^T(x_s - x)\right] \text{ by unbiased gradient and convexity of } f \\
&= \frac{1}{t}\mathbb{E}\left[\sum_{s=1}^{t}\tilde{g}(x_s)^T(x_s - x)\right] \text{ by tower property of conditional expectation}
\end{aligned} \tag{1}$$

Thus we see it now suffices to bound the summand.

$$\begin{aligned}
\tilde{g}_s^T(x_s - x) &= \frac{1}{\eta}\left(\nabla\phi(x_s) - \nabla\phi(y_{s+1})\right)^T(x_s - x) \text{ by the mirror update} \\
&= \frac{1}{\eta}\left(D_\phi(x, x_s) + D_\phi(x_s, y_{s+1}) - D_\phi(x, y_{s+1})\right) \text{ by Bregman divergence property} \\
&= \frac{1}{\eta}\left(D_\phi(x, x_s) + D_\phi(x_s, y_{s+1}) - D_\phi(x, x_{s+1}) - D_\phi(x_{s+1}, y_{s+1}\right) \text{ by Bregman Pythagoras} \\
&= \frac{1}{\eta}\left(\left[D_\phi(x, x_s) - D_\phi(x, x_{s+1})\right] + \left[D_\phi(x_s, y_{s+1}) - D_\phi(x_{s+1}, y_{s+1})\right]\right)
\end{aligned} \tag{2}$$

After we sum over s, the first term will telescope, so we need to bound the second term. By writing out definitions, we have

$$\begin{aligned}
D_\phi(x_s, y_{s+1}) - D_\phi(x_{s+1}, y_{s+1} &= \phi(x_s) - \phi(x_{s+1}) - \nabla\phi(y_{s+1})^T(x_s - x_{s+1}) \\
&\leq (\nabla\phi(x_s) - \nabla\phi(y_{s+1}))^T(x_s - x_{s+1}) - \frac{1}{2}\|x_s - x_{s+1}\|^2 \text{ by strong convexity} \\
&= \eta\tilde{g}_s^T(x_s - x_{s+1}) - \frac{1}{2}\|x_s - x_{s+1}\|^2 \text{ by mirror update} \\
&\leq \eta\|\tilde{g}_s\|_*\|x_s - x_{s+1}\| - \frac{1}{2}\|x_s - x_{s+1}\|^2 \text{ by Holder inequality} \\
&\leq \frac{\eta^2\|\tilde{g}_s\|_*^2}{2} \text{ by } az - bz^2 \leq \frac{a^2}{4b}
\end{aligned} \tag{3}$$

Now, we sum and put expectations back in:

$$\begin{aligned}
\frac{1}{t}\mathbb{E}\left[\sum_{s=1}^{t}\tilde{g}(x_s)^T(x_s-x)\right] &\leq \frac{1}{t}\mathbb{E}\left[\frac{1}{\eta}\sum_{s=1}^{t}D_\phi(x,x_s)-D_\phi(x,x_{s+1})+\frac{1}{\eta}\sum_{s=1}^{t}D_\phi(x_s,y_{s+1})-D_\phi(x_{s+1},y_{s+1})\right] \\
&\leq \frac{1}{t}\mathbb{E}\left[\frac{D_\phi(x,x_1)-D_\phi(x,x_{t+1})}{\eta}+\frac{\eta}{2}\sum_{s=1}^{t}\|\tilde{g}_s\|_*^2\right] \\
&\leq \frac{1}{t}\mathbb{E}\left[\frac{\phi(x_1)-\phi(x_{t+1})}{\eta}+\frac{\eta}{2}\sum_{s=1}^{t}\|\tilde{g}_s\|_*^2\right] \text{ by def of Bregman divergence} \\
&\leq \frac{R^2}{\eta t}+\frac{\eta}{2}\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\|\tilde{g}_s\|_*^2\right] \text{ by def of } R \text{ and linearity of expectation} \\
&\leq \frac{R^2}{\eta t}+\frac{\eta B^2}{2} \text{ by bounded expected gradient norm}
\end{aligned} \tag{4}$$

Therefore, choosing $\eta = \frac{R}{B}\sqrt{\frac{2}{t}}$ gives a convergence rate of

$$\mathbb{E}\left[f(\frac{1}{t}\sum_{s=1}^{t}x_s)\right]-\min_{x\in\mathcal{X}}f(x)\leq RB\sqrt{\frac{2}{t}}$$

which is essentially the same rate as in the non-stochastic case. □

We note that in the Euclidean ($\phi(x)=\frac{1}{2}\|x\|_2^2$), strongly convex setting, stochastic gradient descent with time varying η gives a $1/t$ rate matching the non-stochastic case in the same setting.

3 Smooth stochastic optimization

It turns out that smoothness does not accelerate the rate in the stochastic setting for a general stochastic oracle as it does in the non-stochastic setting (going from $1/\sqrt{t} \to 1/t$ with gradient descent, and even $1/t^2$ using Nesterov's accelerated gradient descent).

However, it is possible to modify SGD in the smooth case to get some useful results. We first prove that under smoothness assumptions, it is possible to interpolate between $1/\sqrt{t}$ and $1/t$ rates dependending on the variance and smoothness parameters.

The following theorem makes this claim precise:

Theorem 3.1. *Let ϕ be a mirror map 1-strongly convex on $\mathcal{X}\cap\mathcal{D}$ w.r.t. $\|\cdot\|$ and let $R^2=\sup_{x\in\mathcal{X}\cap\mathcal{D}}\phi(x)-\phi(x_1)$. Let f be convex and β-smooth w.r.t. $\|\cdot\|$. Further assume that the stochastic oracle satisfies $\mathbb{E}\left[\|\nabla f(x)-\tilde{g}(x)\|_*^2\right]\leq\sigma^2$. Then stochastic mirror descent*

with stepsize $\eta = \frac{1}{\beta+1/\eta}$ *and* $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{t}}$ *satisfies*

$$\mathbb{E}\left[f(\frac{1}{t}\sum_{s=1}^{t} x_{s+1})\right] - f(x^*) \leq R\sigma\sqrt{\frac{2}{t}} + \frac{\beta R^2}{t}$$

Proof. First we show that it suffices to bound $f(x_{s+1}) - f(x^*)$.

$$\mathbb{E}\left[f(\frac{1}{t}\sum_{s=1}^{t} x_{s+1})\right] - f(x^*) \leq \frac{1}{t}\mathbb{E}\left[\sum_{s=1}^{t}(f(x_{s+1}) - f(x^*))\right] \text{ by convexity of } f \tag{5}$$

Since we will telescope later on, we now bound $f(x_{s+1}) - f(x_s)$:

$$\begin{aligned}
f(x_{s+1}) - f(x_s) &\leq \nabla f(x_s)^T(x_{s+1} - x_s) + \frac{\beta}{2}\|x_{s+1} - x_s\|^2 \text{ by } \beta - \text{smoothness} \\
&= \tilde{g}_s^T(x_{s+1} - x_s) + (\nabla f(x_s) - \tilde{g}_s)^T(x_{s+1} - x_s) + \frac{\beta}{2}\|x_{s+1} - x_s\|^2 \\
&\leq \tilde{g}_s^T(x_{s+1} - x_s) + \frac{1}{2}\left(2\|\nabla f(x_s) - \tilde{g}_s\|_*\|x_{s+1} - x_s\|\right) + \frac{\beta}{2}\|x_{s+1} - x_s\|^2 \text{ by Holder} \\
&\leq \tilde{g}_s^T(x_{s+1} - x_s) + \frac{1}{2}\left(\eta\|\nabla f(x_s) - \tilde{g}_s\|_*^2 + \frac{1}{\eta}\|x_{s+1} - x_s\|^2\right) + \frac{\beta}{2}\|x_{s+1} - x_s\|^2 \\
&= \tilde{g}_s^T(x_{s+1} - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 + \frac{1}{2}\left(\frac{1}{\eta} + \beta\right)\|x_{s+1} - x_s\|^2
\end{aligned} \tag{6}$$

where we used the inequality $2ab \leq \eta a^2 + b^2/\eta, \eta > 0$, which follows from AM-GM with $x = \eta a^2, y = b^2/\eta$.

Then, recalling the definition of α-strong convexity $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\alpha}{2}\|y - x\|^2$ and the definition of Bregman divergence $D_f(x, y) = f(x) - f(y) + \nabla f(y)^T(y - x)$, we see that

$$D_f(y, x) = f(y) - f(x) + \nabla f(x)^T(x - y) \geq \frac{\alpha}{2}\|y - x\|^2$$

Thus, since ϕ is 1-strongly convex, we have

$$D_\phi(x_{s+1}, x_s) \geq \frac{1}{2}\|x_{s+1} - x_s\|^2$$

and thus

$$f(x_{s+1}) - f(x_s) \leq \tilde{g}_s^T(x_{s+1} - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 + \left(\frac{1}{\eta} + \beta\right) D_\phi(x_{s+1}, x_s)$$

If we let the stepsize be $\frac{1}{\beta+1/\eta}$, from the mirror descent update, we have

$$\nabla\phi(x_s) - \nabla\phi(y_{s+1}) = \frac{1}{\beta + 1/\eta}\tilde{g}_s$$

Since we know that $x_{s+1} = \Pi^{\phi}_{\mathcal{X}}(y_{s+1})$, and we also have from the Bregman Pythagorean theorem that for any $x \in \mathcal{D} \cap \mathcal{X}$,

$$(\nabla\phi(\Pi^{\phi}_{\mathcal{X}}(y_{s+1})) - \nabla\phi(y_{s+1}))^T(\Pi^{\phi}_{\mathcal{X}}(y_{s+1}) - x) \leq 0$$

Choosing x as the optimal point x^*,

$$(\nabla\phi(x_{s+1}) - \nabla\phi(y_{s+1}))^T(x_{s+1} - x^*) \leq 0$$
$$\nabla\phi(x_{s+1})^T(x_{s+1} - x^*) - \nabla\phi(y_{s+1})^T(x_{s+1} - x^*) + \nabla\phi(x_s)^T(x_{s+1} - x^*) \leq \nabla\phi(x_s)^T(x_{s+1} - x^*) \tag{7}$$

Therefore,

$$\frac{1}{\beta + 1/\eta}\tilde{g}_s^T(x_{s+1} - x^*) \leq (\nabla\phi(x_s) - \nabla\phi(x_{s+1}))^T(x_{s+1} - x^*) = D_\phi(x^*, x_s) - D_\phi(x^*, x_{s+1}) - D_\phi(x_{s+1}, x_s)$$

Thus, we can bound

$$\begin{aligned} f(x_{s+1}) &\leq f(x_s) + \tilde{g}_s^T(x_{s+1} - x^* + x^* - x_s) + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 + \left(\frac{1}{\eta} + \beta\right) D_\phi(x_{s+1}, x_s) \\ &\leq f(x_s) + \tilde{g}_s^T(x^* - x_s) + (\beta + 1/\eta)\left[D_\phi(x^*, x_s) - D_\phi(x^*, x_{s+1})\right] + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2 \end{aligned} \tag{8}$$

Then, by convexity of f, we have $f(x_s) \leq f(x^*) + \nabla f(x_s)^T(x_s - x^*)$, so

$$f(x_{s+1}) - f(x^*) \leq (\tilde{g}_s - \nabla f(x_s))^T(x^* - x_s) + (\beta + 1/\eta)\left[D_\phi(x^*, x_s) - D_\phi(x^*, x_{s+1})\right] + \frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2$$

Then, taking the expectation of the sum over s, we get

$$\begin{aligned} \mathbb{E}\left[f(\frac{1}{t}\sum_{s=1}^{t} x_{s+1})\right] - f(x^*) &\leq \frac{1}{t}\mathbb{E}\left[\sum_{s=1}^{t}(f(x_{s+1}) - f(x^*))\right] \\ &\leq \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[(\tilde{g}_s - \nabla f(x_s))^T(x^* - x_s)\right] \\ &+ \frac{1}{t}\sum_{s=1}^{t}(\beta + 1/\eta)\,\mathbb{E}\left[D_\phi(x^*, x_s) - D_\phi(x^*, x_{s+1})\right] + \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\frac{\eta}{2}\|\nabla f(x_s) - \tilde{g}_s\|_*^2\right] \\ &= \frac{(\beta + 1/\eta)}{t}\mathbb{E}\left[\sum_{s=1}^{t} D_\phi(x^*, x_s) - D_\phi(x^*, x_{s+1})\right] + \frac{\eta\sigma^2}{2} \\ &= \frac{(\beta + 1/\eta)}{t}\mathbb{E}\left[\phi(x_1) - \phi(x_{t+1})\right] + \frac{\eta\sigma^2}{2} \leq \frac{R^2(\beta + 1/\eta)}{t} + \frac{\eta\sigma^2}{2} \end{aligned} \tag{9}$$

since the stochastic oracle has an unbiased gradient, and the variance is bounded by σ^2. Choosing $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{t}}$ gives the desired result.

□

3.1 Minibatch SGD

Now that we have a tunable convergence rate in terms of the standard deviation of the gradient oracle (as defined by $\mathbb{E}\left[\|\tilde{g}_s(x) - \nabla f(x)\|_*^2\right]) \leq \sigma^2$), it makes sense to consider various approaches to reduce the variance in the stochastic case. In the Euclidean setting, the simplest of these is the notion of a minibatch.

When we do minibatch SGD with minibatch size m, we perform the following update:

$$x_{t+1} = \Pi_{\mathcal{X}}\left(x_t - \eta \frac{1}{m}\sum_{i=1}^{m} \tilde{g}_i(x_t)\right)$$

That is, we average over m random indices (m random variables $\tilde{g}_i(x_t)$, conditioned on x_t). In the Euclidean case (where mirror descent and gradient descent are the same thing), assuming that f is β-smooth and $\|\tilde{g}(x)\|_2 \leq B$, we can apply the previous theorem to get a convergence rate for minibatch SGD. First, we calculate the variance bound:

$$\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^{m}\tilde{g}_i(x) - \nabla f(x)\|_2^2\right] = \frac{1}{m}\|\tilde{g}_1(x) - \nabla f(x)\|_2^2 \leq \frac{2B^2}{m} \tag{10}$$

Thus, for larger m, we have a smaller variance and a better convergence rate. However, we need to remember that if we increase m, the cost per iteration increases. Thus, in terms of t calls to the original oracle (t/m minibatch calls), the convergence rate becomes

$$R\sqrt{\frac{2B^2}{m}}\sqrt{\frac{2}{t/m}} + \frac{\beta R^2}{t/m} = 2\frac{RB}{\sqrt{t}} + \frac{m\beta R^2}{t}$$

Thus, for $m \leq \frac{B}{R\beta}\sqrt{t}$, one can obtain a rate which decays as $\frac{3RB}{\sqrt{t}}$. If parallel computation is possible (i.e. there are m processors which can calculate the gradients simultaneously), minibatch SGD is a useful speedup. If many gradients re-occur and can be quickly calculated and re-used, then minibatch SGD is also a useful method.

4 SVRG: Stochastic Variance-Reduced Gradient Descent

The idea of lowering the variance of the gradient calculation without paying too much extra computation is a useful one. In this section we discuss the SVRG algorithm, which applies to sums of smooth and strongly convex functions ($f = \sum_{i=1}^{m} f_i$, where the f_i are β-smooth convex functions, f is α-strongly convex, and m is the size of the dataset). The main idea of SVRG is to center the output of the stochastic oracle in order to reduce the variance. The result is a simple modification of gradient descent: Instead of making gradient updates, we make updates with the following proxy gradient:

$$\nabla f_i(x) - \nabla f_i(y) + \nabla f(y)$$

where y is a sequence of points which also gets updated, a "centering sequence". Note that in expectation, the value of this term is $\nabla f(x)$, so this is also an unbiased estimator. Also note that $\nabla f(y)$ is the full gradient at y, and requires m gradient computations. It turns out that this modification to the gradient can *reduce the variance* of the gradient estimator $\nabla f_i(x)$. One way to think about this is in terms of visualizing a correction to the ∇f_i estimate: The algorithm maintains a point y at which it knows the full gradient and also calculates $\nabla f_i(y)$ at a given iteration. The difference $\nabla f(y) - \nabla f_i(y)$ gives a "correction vector" which the algorithm hopes applies at the point x. This will in fact be the case if x and y are close enough together, which is what happens if y depends on an average x value as the values of x and y approach the optimum. Intuitively, these modifications reduce the variance since as x and y get close to the optimum, $\nabla f_i(x) - \nabla f_i(y)$ will have increasingly smaller variance, and the value of $\nabla f(y)$ will also be small. We now prove this.

Lemma 4.1. *Let $f_1, \cdots, f_m$ be β-smooth convex functions and let $i \in [m]$ be uniformly distributed. Then*

$$\mathbb{E}\left[\|\nabla f_i(x) - \nabla f_i(x^*)\|_2^2\right] \leq 2\beta(f(x) - f(x^*))$$

Proof. Let $g_i(x) = f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x - x^*)$. By convexity of f_i, $g_i(x) \geq 0$ for any x. Recall that for β-smooth functions, we have

$$0 \leq f(a) - f(b) - \nabla f(b)^T(a - b) \leq \frac{\beta}{2}\|a - b\|^2$$

Picking $a = x - \frac{1}{\beta}\nabla f(x), b = x$, we get

$$\begin{aligned} f(a) - f(x) + \frac{1}{\beta}\|\nabla f(x)\|^2 &\leq \frac{1}{2\beta}\|\nabla f(x)\|^2 \\ f(a) - f(x) &\leq -\frac{1}{2\beta}\|\nabla f(x)\|^2 \end{aligned} \tag{11}$$

Therefore, taking $f = g_i$ and noting that $g_i \geq 0$, we must have that $g_i(a) \geq 0$ and therefore $-g_i(x) \leq -\frac{1}{2\beta}\|\nabla g_i(x)\|^2$. This inequality is equivalent to

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|_2^2 \leq 2\beta(f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x - x^*))$$

Since in expectation $\mathbb{E}\left[\nabla f_i(x^*)\right] = 0$ (the optimal point, and ∇f_i is an unbiased estimator), we get the desired result.

□

Now that the general method is somewhat justified, we have another concern to handle: Currently, $\nabla f(y)$ requires m computations. This is expensive, so we need to find a way to sparsely update y. We give the following epochal strategy, where we update y once per epoch.

Let $y^{(1)}$ be an arbitrary initial point. For epoch $s = 1, 2, \cdots$, let the starting $x_1^{(s)}$ for each epoch be the same as $y^{(s)}$. Then, for the k iterations in each epoch, $t = 1, \cdots, k$, update:

Definition 4.2. SVRG Update.

$$x_{t+1}^{(s)} = x_t^{(s)} - \eta \left(\nabla f_{i_t^{(s)}}(x_t^{(s)}) - \nabla f_{i_t^{(s)}}(y^{(s)}) + \nabla f(y^{(s)}) \right)$$

where $i_t^{(s)}$ is drawn uniformly at random and independently from $[m]$. Then, update y:

$$y^{(s+1)} = \frac{1}{k} \sum_{t=1}^{k} x_t^{(s)}$$

Essentially, $y^{(s)}$ for epoch s is the average point from the previous epoch.

Now we prove a convergence rate for SVRG.

Theorem 4.3. *Let $f_1, \cdots, f_m$ be β-smooth convex functions and let f be α-strongly convex. Then SVRG with $\eta = 1/10\beta$ and $k = 20\kappa$ ($\kappa = \beta/\alpha$, the condition number) satisfies*

$$\mathbb{E}\left[f(y^{(s+1)})\right] - f(x^*) \leq 0.9^s (f(y^{(1)}) - f(x^*))$$

Proof. We fix an s, we will show that between epoch steps $y^{(s)}$ and $y^{(s+1)}$, the difference in function values between the candidate point and the optimum decreases by a factor of 0.9. In other words,

$$\mathbb{E}\left[f(y^{(s+1)})\right] - f(x^*) = \mathbb{E}\left[f\left(\frac{1}{k} \sum_{t=1}^{k} x_t^{(s)} \right) \right] - f(x^*) \leq 0.9(f(y^{(s)}) - f(x^*))$$

This implies a geometric convergence rate as stated in the theorem. Simplifying, we only need prove

$$\mathbb{E}\left[f\left(\frac{1}{k} \sum_{t=1}^{k} x_t \right) \right] - f(x^*) \leq 0.9(f(y) - f(x^*))$$

Now consider that $\|x_{t+1} - x^*\|_2^2 = \|x_t - x^*\|_2^2 - 2\eta v_t^T (x_t - x^*) + \eta^2 \|v_t\|_2^2$ where $v_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(y) + \nabla f(y)$, the SVRG update.

Now we bound the last term using our previous lemma

$$\begin{aligned} \mathbb{E}_{i_t}\left[\|v_t\|_2^2\right] &\leq 2\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\|_2^2\right] + 2\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(y) - \nabla f_{i_t}(x^*) - \nabla f(y)\|_2^2\right] \\ &\leq 2\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\|_2^2\right] + 2\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(y) - \nabla f_{i_t}(x^*)\right] \\ &\leq 4\beta(f(x_t) - f(x^*) + f(y) - f(x^*)) \text{ by our lemma} \end{aligned} \tag{12}$$

where the first step follows from $\mathbb{E}_{i_t}\left[\nabla f_{i_t}(x^*)\right] = 0$ and where the second step follows from $\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|_2^2\right] \leq \mathbb{E}\left[\|X\|_2^2\right]$.

Now we bound the middle term

$$-\mathbb{E}_{i_t}\left[v_t^T (x_t - x^*)\right] = -\nabla f(x_t)^T (x_t - x^*) \leq -(f(x_t) - f(x^*))$$

Thus, plugging these two inequalities in, we get

$$\begin{aligned}\mathbb{E}_{i_t}\left[\|x_{t+1}-x^*\|_2^2\right] \leq \|x_t - x^*\|_2^2 - 2\eta(1-2\beta\eta)(f(x_t)-f(x^*)) \\ + 4\beta\eta^2(f(y)-f(x^*))\end{aligned} \tag{13}$$

Now we sum over $t = 1, \cdots, k$ to get

$$\sum_{t=1}^{k}\mathbb{E}_{i_t}\left[\|x_{t+1}-x^*\|_2^2\right] \leq \sum_{t=1}^{k}\|x_t-x^*\|_2^2 - 2\eta(1-2\beta\eta)\sum_{t=1}^{k}(f(x_t)-f(x^*)) + 4\beta\eta^2 k(f(y)-f(x^*))$$

$$\sum_{t=1}^{k}\|x_{t+1}-x^*\|_2^2 - \|x_t-x^*\|_2^2 \leq -2\eta(1-2\beta\eta)\sum_{t=1}^{k}(f(x_t)-f(x^*)) + 4\beta\eta^2 k(f(y)-f(x^*))$$

$$\mathbb{E}\left[\|x_{k+1}-x^*\|_2^2\right] \leq \mathbb{E}\left[\|x_1-x^*\|_2^2\right] - 2\eta(1-2\beta\eta)\mathbb{E}\left[\sum_{t=1}^{k}(f(x_t)-f(x^*))\right] + 4\beta\eta^2 k(f(y)-f(x \tag{14}$$

Since $x_1 = y$ and by α-strong convexity $f(x) - f(x^*) \geq \frac{\alpha}{2}\|x-x^*\|_2^2$, we get

$$\mathbb{E}\left[f\left(\frac{1}{k}\sum_{t=1}^{k}x_t\right)\right] - f(x^*) \leq \left(\frac{1}{\alpha\eta(1-2\beta\eta)k} + \frac{2\beta\eta}{1-2\beta\eta}\right)(f(y)-f(x^*))$$

Choosing $\eta = 1/10\beta$ and $k = 20\kappa$ gives the desired result.

□

5 Random Coordinate Descent

Another way a stochastic gradient estimator can come into play is if we don't use full gradient information at each step. Suppose that instead of calculating the full gradient at a point, we only calculated the gradient along a specific coordinate. In other words, for dimension n,

$$\nabla f(x) = \sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(x)e_j$$

involves calculating n real numbers and combining them into a full gradient value at a point. Note that here, the function f could be $\sum_{i=1}^{m} f_i$, or it could be a single f_i. The stochasticity in this example will come from a random variable over dimensions, not over samples.

We define Random Coordinate Descent (RCD) as the following algorithm:

Definition 5.1. Random Coordinate Descent (RCD).
Start with an arbitrary initial point $x_1 \in \mathbb{R}^n$. Then, we update

$$x_{s+1} = x_s - \eta\nabla_{i_s}f(x)e_{i_s}$$

where i_s is drawn uniformly at random from $[n]$ and $\nabla_j f : \mathbb{R}^n \to \mathbb{R}$ is taken to mean $\frac{\partial f}{\partial x_j}$.

RCD can therefore be viewed as stochastic gradient descent where the stochasticity is NOT over the samples, as previously, but is instead over the dimensions. Specifically, the gradient oracle for RCD viewed as SGD is

$$\tilde{g}(x) = n\nabla_j f(x) e_j$$

where j is drawn uniformly at random from $[n]$. We immediately see that this oracle is unbiased (this is the purpose of multiplying by n) and we can furthermore calculate

$$\mathbb{E}\left[\|\tilde{g}(x)\|_2^2\right] = \frac{1}{n}\sum_{j=1}^{n} \|n\nabla_j f(x) e_j\|_2^2 = n\|\nabla f(x)\|_2^2$$

Therefore, in the Euclidean case, the first theorem presented in this exposition immediately gives the following theorem taking $B = L\sqrt{n}$:

Theorem 5.2. *Let f be convex and L-Lipschitz on $\mathbb{R}^n$, then RCD with $\eta = \frac{R}{L}\sqrt{\frac{2}{nt}}$ gives*

$$\mathbb{E}\left[f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right)\right] - \min_{x\in\mathcal{X}} f(x) \leq RL\sqrt{\frac{2n}{t}}$$

As we can see, this vanilla version of RCD seems somewhat useless as it requires n times more iterations than gradient descent would to obtain the same accuracy. This makes sense, because we essentially need to sample a factor of n more times in each direction in order to get the same expected amount of information, as in the problem as it stands there is no information gained about the other coordinates when we sample from one coordinate.

We can fix this issue by introducing the notion of *directional smoothness*.

Definition 5.3. Directional smoothness.
For a function f, there exist $\beta_1, \cdots, \beta_n$ such that for any $i \in [n]$, $x \in \mathbb{R}^n, u \in \mathbb{R}$, we have

$$|\nabla_i f(x + ue_i) - \nabla_i f(x)| \leq \beta_i |u|$$

If f is twice differentiable, this statement amounts to bounding the diagonal of the Hessian matrix: $(\nabla^2 f(x))_{i,i} \leq \beta_i$. In particular, since the smoothness of a function is an upper bound on the maximal eigenvalue of the Hessian, and the maximum eigenvalue is bounded by the trace, we have that f is β-smooth where $\beta \leq \sum_{i=1}^n \beta_i$.

The main idea we can use to modify RCD is now we use information about the smoothness in each direction when taking gradient updates. Gradient steps in a given direction will now be inversely proportional to the smoothness in that direction:

$$x_{s+1} = x_s - \frac{1}{\beta_{i_s}}\nabla_{i_s} f(x) e_{i_s}$$

where i_s is as usual drawn independently. It turns out to be furthermore useful to take steps according to a distribution p_γ, defined over the coordinates $i \in [n]$ with $\gamma \geq 0$:

$$p_\gamma(i) = \frac{\beta_i^\gamma}{\sum_{j=1}^n \beta_j^\gamma}$$

It turns out (Nesterov [2012]) that the rate of convergence for modified RCD can be expressed in terms of the dual norms $\|\cdot\|_{[\gamma]}$ and $\|\cdot\|_{[\gamma]}^*$, defined by

Definition 5.4. $\|\cdot\|_{[\gamma]}$ and $\|\cdot\|_{[\gamma]}^*$.

$$\begin{aligned} \|x\|_{[\gamma]} &= \sqrt{\sum_{i=1}^n \beta_i^\gamma x_i^2} \\ \|x\|_{[\gamma]}^* &= \sqrt{\sum_{i=1}^n \frac{1}{\beta_i^\gamma} x_i^2} \end{aligned} \tag{15}$$

Then we have the following theorem:

Theorem 5.5. *Let f be convex s.t. for $u \in \mathbb{R}$, $f(x + ue_i)$ is β_i-smooth for any $i \in [n]$, $x \in \mathbb{R}^n$. Then RCD(γ) satisfies for $t \geq 2$*

$$\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{2R_{1-\gamma}(x_1)^2 \sum_{i=1}^n \beta_i^\gamma}{t-1}$$

where

$$R_{1-\gamma}(x_1) = \sup_{x \in \mathbb{R}^n : f(x) \leq f(x_1)} \|x - x^*\|_{[1-\gamma]}$$

Standard gradient descent attains a rate of $\beta\|x_1 - x^*\|_2^2/t$ where $\beta \leq \sum_{i=1}^n \beta_i$. Therefore, for $\gamma = 1$, RCD is a great improvement over gradient descent: The same number of iterations will give similar accuracy, but the computational cost of each RCD step is much less.

Proof. As we saw in the SVRG lemma, we have

$$f(x - \frac{1}{\beta_i}\nabla_i f(x) e_i) - f(x) \leq -\frac{1}{2\beta_i}(\nabla_i f(x))^2$$

Therefore, we can bound

$$\begin{aligned} \mathbb{E}_{i_s}\left[f(x_{s+1}) - f(x_s)\right] &= \sum_{i=1}^n p_\gamma(i) \left(f(x - \frac{1}{\beta_i}\nabla_i f(x) e_i) - f(x) \right) \\ &\leq -\sum_{i=1}^n \frac{p_\gamma(i)}{2\beta_i}(\nabla_i f(x_s))^2 \\ &= -\frac{1}{2\sum_{i=1}^n \beta_i^\gamma} \left(\|\nabla f(x_s)\|_{[1-\gamma]}^* \right)^2 \end{aligned} \tag{16}$$

Now denote $\delta_s = \mathbb{E}\left[f(x_s)\right] - f(x^*)$. Note that the above calculation implies $f(x_{s+1}) \leq f(x_s)$ since the RHS is non-positive. Therefore, by definition of $R_{1-\gamma}(x_1)$,

$$\begin{aligned}\delta_s = \mathbb{E}\left[f(x_s) - f(x^*)\right] &\leq \nabla f(x_s)^T(x_s - x^*) \text{ by convexity} \\ &\leq \|x_s - x^*\|_{[1-\gamma]}\|\nabla f(x_s)\|^*_{[1-\gamma]} \text{ by Holder} \\ &\leq R_{1-\gamma}(x_1)\|\nabla f(x_s)\|^*_{[1-\gamma]}\end{aligned} \tag{17}$$

Thus, we have

$$-\left(\|\nabla f(x_s)\|^*_{[1-\gamma]}\right)^2 \leq -\left(\frac{\delta_s}{R_{1-\gamma}(x_1)}\right)^2$$

Now, we see

$$\begin{aligned}\delta_{s+1} - \delta_s &= \mathbb{E}\left[f(x_{s+1})\right] - f(x^*) - \mathbb{E}\left[f(x_s)\right] + f(x^*) \\ \delta_{s+1} &\leq \delta_s - \frac{1}{2\sum_{i=1}^n \beta_i^\gamma}\left(\|\nabla f(x_s)\|^*_{[1-\gamma]}\right)^2 \\ &\leq \delta_s - \frac{1}{2R^2_{1-\gamma}(x_1)\sum_{i=1}^n \beta_i^\gamma}\delta_s^2\end{aligned} \tag{18}$$

Let $\omega = \frac{1}{2R^2_{1-\gamma}(x_1)\sum_{i=1}^n \beta_i^\gamma}$. Then, we have

$$\begin{aligned}\omega\delta_s^2 + \delta_{s+1} &\leq \delta_s \\ \omega\frac{\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} &\leq \frac{1}{\delta_{s+1}} \\ \omega &\leq \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \text{ since } \delta_s \geq \delta_{s+1} \\ \sum_{s=1}^{t-1}\omega &\leq \sum_{s=1}^{t-1}\frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \\ \omega(t-1) &\leq \frac{1}{\delta_t} - \frac{1}{\delta_1} \leq \frac{1}{\delta_t} \\ \delta_t &\leq \frac{2R^2_{1-\gamma}(x_1)\sum_{i=1}^n \beta_i^\gamma}{t-1} \\ \mathbb{E}\left[f(x_t)\right] - f(x^*) &\leq \frac{2R^2_{1-\gamma}(x_1)\sum_{i=1}^n \beta_i^\gamma}{t-1}\end{aligned} \tag{19}$$

as desired. □

Now if in addition to directional smoothness one assumes strong convexity, RCD attains an even faster rate.

Theorem 5.6. *Let $\gamma \geq 0$. Let f be α-strongly convex w.r.t. $\|\cdot\|_{[1-\gamma]}$ s.t. for $u \in \mathbb{R}$ $f(x+ue_i)$ is β_i-smooth for any $i \in [n]$, $x \in \mathbb{R}^n$. Let $\kappa_\gamma = \frac{1}{\alpha}\sum_{i=1}^n \beta_i^\gamma$, then RCD($\gamma$) satisfies*

$$\mathbb{E}\left[f(x_{t+1})\right] - f(x^*) \leq \left(1 - \frac{1}{\kappa_\gamma}\right)^t \left(f(x_1) - f(x^*)\right)$$

First we prove a simple lemma:

Lemma 5.7. *Let f be α-strongly convex w.r.t. $\|\cdot\|$ on $\mathbb{R}^n$, then*

$$f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|_*^2$$

Proof. Applying strong convexity, Holder, and the fact that $bz - az^2 \leq \frac{b^2}{4a}$, we have

$$\begin{aligned} f(x) - f(y) &\leq \nabla f(x)^T(x-y) - \frac{\alpha}{2}\|x-y\|_2^2 \\ &\leq \|\nabla f(x)\|_*\|x-y\| - \frac{\alpha}{2}\|x-y\|_2^2 \\ &\leq \frac{1}{2\alpha}\|\nabla f(x)\|_*^2 \end{aligned} \tag{20}$$

and we finish by taking $y = x^*$. □

Now we prove the theorem.

Proof. Note that the lemma implies that

$$-\left(\|\nabla f(x_s)\|_{[1-\gamma]}^*\right)^2 \leq -2\alpha\delta_s$$

In the previous case, we had shown

$$\delta_{s+1} \leq \delta_s - \frac{1}{2\sum_{i=1}^n \beta_i^\gamma}\left(\|\nabla f(x_s)\|_{[1-\gamma]}^*\right)^2$$

Therefore,

$$\delta_{s+1} \leq \left(1 - \frac{\alpha}{\sum_{i=1}^n \beta_i^\gamma}\right)\delta_s$$

Fixing $s = 1$ and applying t times, we get

$$\delta_{t+1} \leq \left(1 - \frac{1}{\kappa_\gamma}\right)^t \delta_1$$

which is precisely the desired result. □