1 Outline

In this lecture, we study

- coordinate descent,
- random coordinate descent,
- variance-reduced stochastic methods.

2 Coordinate Descent

When training a machine learning model, we often deal with a huge number of features and parameters. Then the training process corresponds to a high-dimensional optimization problem, in which computing the gradient or its stochastic estimate is expensive. On the other hand, it is often easy to deduce directional derivatives along the coordinate directions. Moreover, some structured optimization problems admit a decomposition with respect to a partition of the coordinates. For example, we have regularizers $f(x) = \|x\|_2^2$ and $f(x) = \|x\|_1$. In addition, regularizers that induce "group sparsity" are proposed, and they are of the form

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

where $S_1 \cup \cdots \cup S_m = [d]$ and $x = (x_{S_1}, \ldots, x_{S_m})$. Coordinate descent is a widely used optimization method that runs with directional derivatives and thus provides an efficient framework for tacking the abovementioned applications.

For $i \in [d]$, let $\partial_i f(x)$ denote the directional derivative of f at x along the coordinate direction $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^d$:

$$\partial_i f(x) = \lim_{\delta \to 0} \frac{f(x + \delta e_i) - f(x)}{\delta}.$$

At each iteration t, coordinate descent takes an index $i_t \in [d]$ and deduce the next iterate x_{t+1} from the current solution x_t based on

$$x_{t+1} = x_t - \eta_t \partial_{i_t} f(x_t) e_{i_t}.$$

Basically, coordinate descent updates one coordinate at a time. There are many strategies for choosing an index at each iteration. In this section, we consider random sampling-based coordinate descent implementations.

The most basic version is to sample a coordinate uniformly at random. In fact, this version is an instance of stochastic gradient descent. To see this, we take

$$g_t = d \cdot \partial_{i_t} f(x_t) e_{i_t}$$

and note that

$$\mathbb{E}\left[g_t \mid x_t\right] = \sum_{i=1}^{d} \frac{1}{d} \cdot d \cdot \partial_i f(x_t) e_i = \nabla f(x_t).$$

Hence, g_t is an unbiased estimator of $\nabla f(x_t)$, and coordinate descent runs with the update rule

$$x_{t+1} = x_t - \frac{\eta_t}{d} g_t$$

with step size η_t/d. Moreover, we have

$$\mathbb{E}\left[\|g_t\|_2^2 \mid x_t\right] = \sum_{i=1}^{d} \frac{1}{d} \cdot d^2 |\partial_i f(x_t)|^2 = d\|\nabla f(x_t)\|_2^2.$$

Algorithm 1 Coordinate Descent

Initialize $x_1 \in \mathbb{R}^d$.
for $t = 1, \ldots, T$ **do**
  Sample an index $i_t \in [d]$ uniformly at random.
  Update $x_{t+1} = x_t - \eta_t \partial_{i_t} f(x_t) e_{i_t}$ for a step size $\eta_t > 0$.
end for
Return $(1/T)\sum_{t=1}^{T} x_t$.

Theorem 8.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that is L-Lipschitz continuous in the ℓ_2-norm. Then choosing $\eta_t = \sqrt{d/T}$ for $t \geq 1$, we have*

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2 + L^2}{2}\sqrt{\frac{d}{T}}$$

where $x^* \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$ and the expectation is taken over the random choice of coordinates.

Proof. The update rule of coordinate descent implies that

$$g_t^\top (x_t - x^*) \leq \frac{d}{2\eta_t}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\eta_t}{2d}\|g_t\|_2^2.$$

Note that

$$\mathbb{E}\left[g_t^\top (x_t - x^*)\right] = \mathbb{E}\left[\mathbb{E}\left[g_t^\top (x_t - x^*) \mid x_t\right]\right] = \mathbb{E}\left[\nabla f(x_t)^\top (x_t - x^*)\right] \geq \mathbb{E}\left[f(x_t) - f(x^*)\right].$$

Moreover,

$$\mathbb{E}\left[\|g_t\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|g_t\|_2^2 \mid x_t\right]\right] = \mathbb{E}\left[d\|\nabla f(x_t)\|_2^2\right] \leq dL^2.$$

Then it follows that

$$\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{d}{2\eta_t}\left(\mathbb{E}\left[\|x_t - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right]\right) + \frac{\eta_t}{2}L^2.$$

Summing up this inequality for $t = 1, \ldots, T$ and dividing the resulting one by T, we deduce that

$$\mathbb{E}\left[f(x_t)\right] - f(x^*) \leq \frac{d\|x_1 - x^*\|_2^2}{2\eta T} + \frac{\eta}{2}L^2$$

where $\eta = \sqrt{d/T}$. □

3 Random Coordinate Descent

Assume that for $i \in [d]$,

$$|\partial_i f(x + \delta e_i) - \partial_i f(x)| \leq \beta_i |\delta| .$$

Note that this is a coordinate version of smoothness. In fact, if f is β-smooth in the ℓ_2-norm, it follows that

$$|\partial_i f(x + \delta e_i) - \partial_i f(x)| \leq \|\nabla f(x + \delta e_i) - \nabla f(x)\|_2 \leq \beta |\delta| .$$

Then we consider a random index sampling strategy which samples index $i \in [d]$ with probability

$$\frac{\beta_i^\gamma}{\sum_{j=1}^d \beta_j^\gamma}$$

for some $\gamma > 0$. Let $\mathbb{P}(\gamma)$ denote the corresponding probability distribution over the indices. Then we consider coordinate descent with the following update rule. At each iteration t, we sample an index i_t from distribution $\mathbb{P}(\gamma)$ and take

$$x_{t+1} = x_t - \frac{1}{\beta_{i_t}} \partial_{i_t} f(x_t) e_{i_t}.$$

We refer to this version of coordinate descent as **random coordinate descent** and use notation RCD(γ) to specify the parameter γ. Unlike the previous version of coordinate descent, RCD(γ) is

Algorithm 2 RCD(γ)

Initialize $x_1 \in \mathbb{R}^d$.
for $t = 1, \ldots, T$ **do**
  Sample an index $i_t \in [d]$ from the distribution $\mathbb{P}(\gamma)$.
  Update $x_{t+1} = x_t - \frac{1}{\beta_{i_t}} \partial_{i_t} f(x_t) e_{i_t}$ for a step size $\eta_t > 0$.
end for
Return x_{T+1}.

not an instance of SGD. To see this, we consider

$$\mathbb{E}\left[\frac{1}{\beta_{i_t}} \partial_{i_t} f(x_t) e_{i_t}\right] = \sum_{i=1}^n \frac{1}{\sum_{j=1}^d \beta_j^\gamma} \beta_i^{\gamma-1} \partial_i f(x_t) e_i,$$

which explains that the direction

$$g_t = \frac{1}{\beta_{i_t}} \partial_{i_t} f(x_t) e_{i_t}$$

is not an unbiased estimator of the gradient $\nabla f(x_t)$. The next theorem provides a convergence guarantee of RCD(γ).

Theorem 8.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that satisfies $|\partial_i f(x + \delta e_i) - \partial_i f(x)| \leq \beta_i |\delta|$ for $i \in [d]$. Then RCD(γ) guarantees that*

$$\mathbb{E}[f(x_{T+1})] - f(x^*) \leq \frac{2R^2 \sum_{i=1}^d \beta_i^\gamma}{T}$$

where $x^* \in argmin_{x \in \mathbb{R}^d} f(x)$ and

$$R^2 = \sup_{x \in \mathbb{R}^d : f(x) \leq f(x_1)} \sum_{i=1}^d \beta_i^{1-\gamma} (x - x^*)_i^2.$$

Proof. Note that for any $x \in \mathbb{R}^d$,

$$h_{i,x}(\delta) = f(x + \delta e_i)$$

is a convex function that is β_i-smooth. Moreover, we have that

$$\nabla_\delta h_{i,x}(0) = \lim_{\epsilon \to 0} \frac{h_{i,x}(\epsilon) - h_{i,x}(0)}{\epsilon} = \lim_{\epsilon \to 0} \frac{f(x + \epsilon e_i) - f(x)}{\epsilon} = \partial_i f(x).$$

Note that the RCD(γ) step applied at iteration t is equivalent to gradient descent applied to h_{i_t,x_t} which is β_{i_t}-smooth. Based on the analysis of gradient descent for smooth convex minimization, it follows that

$$f\left(x - \frac{1}{\beta_i}\partial_i f(x) e_i\right) - f(x) = h_{i,x}\left(-\frac{1}{\beta_i}\partial_i f(x)\right) - h_{i,x}(0) \leq -\frac{1}{2\beta_i}\|\nabla_\delta h_{i,x}(0)\|_2^2 = -\frac{1}{2\beta_i}\partial_i f(x)^2.$$

This implies that RCD(γ) is a descent method:

$$f(x_1) \geq f(x_2) \geq \cdots \geq f(x_{T+1}).$$

Furthermore, for any fixed x,

$$\begin{aligned}
\mathbb{E}_{i \sim \mathbb{P}(\gamma)}\left[f\left(x - \frac{1}{\beta_i}\partial_i f(x) e_i\right) - f(x)\right] &\leq \mathbb{E}_{i \sim \mathbb{P}(\gamma)}\left[-\frac{1}{2\beta_i}\partial_i f(x)^2\right] \\
&= \sum_{i=1}^d \frac{\beta_i^\gamma}{\sum_{j=1}^d \beta_j^\gamma} \cdot -\frac{1}{2\beta_i}\partial_i f(x)^2 \\
&= -\frac{1}{2\sum_{j=1}^d \beta_j^\gamma} \sum_{i=1}^d \beta_i^{\gamma-1}\partial_i f(x)^2.
\end{aligned}$$

This implies that

$$\mathbb{E}\left[f\left(x_{t+1}\right) - f(x_t) \mid x_t\right] \leq -\frac{1}{2\sum_{j=1}^d \beta_j^\gamma} \sum_{i=1}^d \beta_i^{\gamma-1}\partial_i f(x_t)^2.$$

Moreover, convexity of f implies that

$$\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \\
&\leq \left(\sum_{i=1}^d \beta_i^{\gamma-1}\partial_i f(x_t)^2\right)^{1/2} \left(\sum_{i=1}^d \beta_i^{1-\gamma}(x_t - x^*)_i^2\right)^{1/2} \\
&\leq R\left(\sum_{i=1}^d \beta_i^{\gamma-1}\partial_i f(x_t)^2\right)^{1/2}
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the third inequality is due to the choice of R. Combining the last two inequalitis, it follows that

$$\mathbb{E}\left[f\left(x_{t+1}\right) - f(x_t) \mid x_t\right] \leq -\frac{1}{2R^2\sum_{j=1}^d \beta_j^\gamma}\left(f(x_t) - f(x^*)\right)^2.$$

Taking the expectation, we obtain

$$\mathbb{E}\left[f(x_{t+1}) - f(x^*)\right] - \mathbb{E}\left[f(x_t) - f(x^*)\right] \leq -\frac{1}{2R^2\sum_{j=1}^d \beta_j^\gamma}\mathbb{E}\left[\left(f(x_t) - f(x^*)\right)^2\right].$$

Here, as $f(x_{t+1}) \le f(x_t)$, we have $f(x_{t+1}) - f(x^*) \le f(x_t) - f(X^*)$ and thus

$$\mathbb{E}\left[(f(x_t) - f(x^*))^2\right] \ge \mathbb{E}\left[f(x_t) - f(x^*)\right]^2 \ge \mathbb{E}\left[f(x_{t+1}) - f(x^*)\right] \cdot \mathbb{E}\left[f(x_t) - f(x^*)\right].$$

Therefore, it follows that

$$\frac{1}{\mathbb{E}\left[f(x_t) - f(x^*)\right]} - \frac{1}{\mathbb{E}\left[f(x_{t+1}) - f(x^*)\right]} \le -\frac{1}{2R^2 \sum_{j=1}^d \beta_j^\gamma}.$$

Summing up this inequality for $t \ge 1$, we deduce that

$$\mathbb{E}\left[f(x_{T+1}) - f(x^*)\right] \le \frac{2R^2 \sum_{j=1}^d \beta_j^d}{T},$$

as required. □

4 Variance-Reduced (VR) Stochastic Methods

Recall that stochastic gradient descent guarantees that if the step size is set to

$$\eta_t = \min\left\{\frac{1}{\beta}, \frac{\|x_1 - x^*\|_2}{\sigma\sqrt{2T}}\right\}$$

for $t \ge 1$, we deduce

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=2}^{T+1} x_t\right)\right] - f(x^*) \le \frac{\beta\|x_1 - x^*\|_2^2}{2T} + \frac{\sigma\|x_1 - x^*\|_2\sqrt{2}}{\sqrt{T}}$$

when f is β-smooth. Note that the second term is incurred due to the variance σ^2 of estimating the gradient. Basically, even when the objective function f is smooth, we may have to choose a small step size of order $O(1/\sqrt{T})$. Motivated by this, we develop algorithms that are sample-efficient, and at the same time, recover near-optimal performance guarantees.

We consider

$$\text{minimize}_{x \in \mathbb{R}^d} \quad f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$$

which is called the **finite-sum problem**. In stochastic optimization, we had the objective of

$$\mathbb{E}\left[f(x, \xi)\right].$$

Sampling n random vectors $\xi_1, \ldots, \xi_n$, we obtain n sampled functions $f(x, \xi_1), \ldots, f(x, \xi_n)$. Moreover,

$$\frac{1}{n}\sum_{i=1}^n f(x, \xi_i)$$

is an estimator of the original objective function. Taking $f_i(x) = f(x, \xi_i)$, we get the above optimization problem. Hence, in the context of stochastic optimization, the problem is often called the **empirical risk minimization (ERM)** and the **sample average approximation (SAA)**.

It is widely known that stochastic gradient descent works well for the finite-sum problem. In the previous section, we learned that taking a mini-batch of stochastic gradients can reduce the variance term. In fact, there are other ways of reducing the variance, and they are often called **variance reduced (VR) stochastic methods**. Among many of these methods, we mention a few below.

- Stochastic Average Gradient (SAG) [SLRB17].
- SAGA [DBLJ14].
- Stochastic Variance Reduced Gradient (SVRG) [JZ13].

4.1 Stochastic Variance Reduced Gradient (SVRG)

In particular, we introduce SVRG for this lecture. To elaborate, we select an index r from $\{1, \ldots, n\}$ uniformly at random. Then for any two points x and y, consider

$$\hat{g}_x = \nabla f_r(x) - (\nabla f_r(y) - \nabla f(y)).$$

By the random choice of r, it follows that

$$\begin{aligned}\mathbb{E}\left[\hat{g}_x\right] &= \mathbb{E}\left[\nabla f_r(x)\right] - (\mathbb{E}\left[\nabla f_r(y)\right] - \nabla f(y)) \\ &= \nabla f(x) - (\nabla f(y) - \nabla f(y)) \\ &= \nabla f(x).\end{aligned}$$

In particular, when $y = x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$, we have

$$\hat{g}_x = \nabla f_r(x) - \nabla f_r(x^*).$$

Moreover, we can use

Lemma 8.3. *If $f_1, \ldots, f_n$ are convex and β-smooth in the ℓ_2 norm, then*

$$\mathbb{E}_{r \sim \mathbb{P}}\left[\|\nabla f_r(x) - \nabla f_r(x^*)\|_2^2\right] \leq 2\beta(f(x) - f(x^*))$$

where $\mathbb{P}$ is the uniform distribution over $\{1, \ldots, n\}$ and $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$.

Proof. Note that

$$g_r(x) = f_r(x) - \left(f_r(x^*) + \nabla f_r(x^*)^\top (x - x^*)\right) \geq 0$$

because f_r is convex. Moreover, f_r is β-smooth, and we have

$$\|\nabla g_r(x) - \nabla g_r(y)\|_2 = \|\nabla f_r(x) - \nabla f_r(x^*) - \nabla f_r(y) + \nabla f_r(x^*)\|_2 = \|\nabla f_r(x) - \nabla f_r(y)\|_2,$$

implying in turn that g_r is β-smooth. Then it follows that

$$g_r\left(x - \frac{1}{\beta}\nabla g_r(x)\right) \leq g_r(x) - \frac{1}{2\beta}\|\nabla g_r(x)\|_2^2.$$

As $g_r \geq 0$, we obtain

$$\|\nabla g_r(x)\|_2^2 \leq 2\beta g_r(x).$$

By the definition of g_r, this is equivalent to the following.

$$\|\nabla f_r(x) - \nabla f_r(x)\|_2 \leq 2\beta\left(f_r(x) - f_r(x^*) - \nabla f_r(x^*)^\top (x - x^*)\right).$$

Taking the expection of each side with respect to r,

$$\begin{aligned}\mathbb{E}\left[\|\nabla f_r(x) - \nabla f_r(x)\|_2\right] &\leq 2\beta\left(\mathbb{E}\left[f_r(x)\right] - \mathbb{E}\left[f_r(x^*)\right] - \mathbb{E}\left[\nabla f_r(x^*)^\top (x - x^*)\right]\right) \\ &= 2\beta\left(f(x) - f(x^*) - \nabla f(x^*)^\top (x - x^*)\right) \\ &= 2\beta\left(f(x) - f(x^*)\right),\end{aligned}$$

as required. $\square$

Algorithm 3 Stochastic variance reduced gradient (SVRG) descent

Initialize $x_1 \in C$.
for $t = 1, \ldots, T$ **do**
  $y_1 = x_t$.
  for $k = 1, \ldots, B$ **do**
    Sample r from $\{1, \ldots, n\}$ uniformly at random.
    Update $y_{k+1} = y_k - \eta(\nabla f_r(y_k) - (\nabla f_r(x_t) - \nabla f(x_t)))$.
  end for
  Update $x_{t+1} = \frac{1}{B}\sum_{k=1}^{B} y_k$.
end for
Return x_{T+1}.

Lemma 8.3 basically bounds the variance term $\mathbb{E}\left[\|\hat{g}_x\|_2^2\right]$ given by $\hat{g}_x = \nabla f_r(x) - \nabla f_r(x^*)$. Based on this result, we consider the following algorithm.

In the inner loop, we obtain a stochastic estimator of the gradient, $\nabla f_r(y_k)$, as in each iteration of SGD. On the other hand, the outer loop requires computing the exact gradient, $\nabla f(x_t)$.

4.2 SVRG analysis

Theorem 8.4. *Assume that $f_1, \ldots, f_n$ are β-smooth and $f = (1/n)\sum_{i=1}^{n} f_i$ is α-strongly convex with respect to the ℓ_2 norm. Setting $\eta = 1/(6\beta)$ and $B = 36\beta/\alpha$, x_{T+1} returned by Algorithm 3 satisfies*

$$\mathbb{E}\left[f(x_{T+1})\right] - f(x^*) \leq \left(\frac{3}{4}\right)^T (f(x_1) - f(x^*))$$

where $x^* \in argmin_{x\in\mathbb{R}^d} f(x)$.

Proof. Let

$$g_k = \nabla f_r(y_k) - \nabla f_r(x_t) + \nabla f(x_t).$$

Note that

$$\begin{aligned}\|y_{k+1} - x^*\|_2^2 &= \|y_k - \eta g_k - x^*\|_2^2 \\ &= \|y_k - x^*\|_2^2 - 2\eta g_k^\top (y_k - x^*) + \eta^2\|g_k\|_2^2.\end{aligned} \tag{8.1}$$

Let us consider the third term $\eta^2\|g_k\|_2^2$ in the right-hand side of (8.1). Note that

$$\begin{aligned}&\mathbb{E}\left[\|g_k\|_2^2 \mid y_k\right] \\ &= \mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x_t) + \nabla f(x_t)\|_2^2 \mid y_k\right] \\ &= \mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x^*) + \nabla f_r(x^*) - \nabla f_r(x_t) + \nabla f(x_t)\|_2^2 \mid y_k\right] \\ &\leq \mathbb{E}\left[2\|\nabla f_r(y_k) - \nabla f_r(x^*)\|_2^2 + 2\| - \nabla f_r(x^*) + \nabla f_r(x_t) - \nabla f(x_t)\|_2^2 \mid y_k\right] \\ &= 2\mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x^*)\|_2^2 \mid y_k\right] + 2\mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t) - \nabla f(x_t)\|_2^2 \mid y_k\right]\end{aligned} \tag{8.2}$$

where the inequality is because $\|a - b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$. Moreover, the second term in the right-hand side of (8.2) can be bounded as follows.

$$
\begin{aligned}
&\mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)-\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 - 2\nabla f(x_t)^\top(\nabla f_r(x_t)-\nabla f_r(x^*)) + \|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - 2\nabla f(x_t)^\top \mathbb{E}\left[\nabla f_r(x_t)-\nabla f_r(x^*) \mid \ y_k\right] \\
&\quad + \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - 2\nabla f(x_t)^\top(\nabla f(x_t)-\nabla f(x^*)) \\
&\quad + \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - 2\nabla f(x_t)^\top \nabla f(x_t) + \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&\leq \mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 \mid \ y_k\right].
\end{aligned} \tag{8.3}
$$

Combining (8.2) and (8.3), it follows that

$$
\begin{aligned}
&\mathbb{E}\left[\|g_k\|_2^2 \mid \ y_k\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f_r(y_k)-\nabla f_r(x^*)\|_2^2 \mid \ y_k\right] + 2\mathbb{E}\left[\|-\nabla f_r(x^*)+\nabla f_r(x_t)\|_2^2 \mid \ y_k\right] \\
&\leq 4\beta(f(y_k)-f(x^*)) + 4\beta(f(x_t)-f(x^*)) \\
&= 4\beta(f(y_k)-f(x^*)+f(x_t)-f(x^*)).
\end{aligned} \tag{8.4}
$$

Applying the tower rule to (8.4),

$$
\begin{aligned}
\mathbb{E}\left[\|g_k\|_2^2 \mid \ x_t\right] &= \mathbb{E}\left[\mathbb{E}\left[\|g_k\|_2^2 \mid \ y_k\right] \mid \ x_t\right] \\
&\leq \mathbb{E}\left[4\beta(f(y_k)-f(x^*)+f(x_t)-f(x^*)) \mid \ x_t\right] \\
&= 4\beta(\mathbb{E}\left[f(y_k) \mid \ x_t\right]-f(x^*)+f(x_t)-f(x^*)).
\end{aligned} \tag{8.5}
$$

Next, we consider the term $-2\eta g_k^\top(y_k-x^*)$ in (8.1).

$$
\begin{aligned}
\mathbb{E}\left[-2\eta g_k^\top(y_k-x^*) \mid \ y_k\right] &= -2\eta \mathbb{E}\left[g_k \mid \ y_k\right]^\top (y_k-x^*) \\
&= -2\eta \mathbb{E}\left[\nabla f_r(y_k)-\nabla f_r(x_t)+\nabla f(x_t) \mid \ y_k\right]^\top (y_k-x^*) \\
&= -2\eta \nabla f(y_k)^\top(y_k-x^*) \\
&\leq -2\eta(f(y_k)-f(x^*)).
\end{aligned} \tag{8.6}
$$

Again, applying the tower rule to (8.6),

$$
\begin{aligned}
\mathbb{E}\left[-2\eta g_k^\top(y_k-x^*) \mid \ x_t\right] &= \mathbb{E}\left[\mathbb{E}\left[-2\eta g_k^\top(y_k-x^*) \mid \ y_k\right] \mid \ x_t\right] \\
&\leq \mathbb{E}\left[-2\eta(f(y_k)-f(x^*)) \mid \ x_t\right] \\
&= -2\eta(\mathbb{E}\left[f(y_k) \mid \ x_t\right]-f(x^*))
\end{aligned} \tag{8.7}
$$

Combining (8.1), (8.5), and (8.7), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|y_{k+1}-x^*\|_2^2 \mid \ x_t\right] &\leq \mathbb{E}\left[\|y_k-x^*\|_2^2 \mid \ x_t\right] - 2\eta(\mathbb{E}\left[f(y_k) \mid \ x_t\right]-f(x^*)) \\
&\quad + 4\eta^2\beta(\mathbb{E}\left[f(y_k) \mid \ x_t\right]-f(x^*)+f(x_t)-f(x^*)) \\
&= \mathbb{E}\left[\|y_k-x^*\|_2^2 \mid \ x_t\right] - 2\eta(1-2\eta\beta)(\mathbb{E}\left[f(y_k) \mid \ x_t\right]-f(x^*)) \\
&\quad + 4\eta^2\beta(f(x_t)-f(x^*))
\end{aligned} \tag{8.8}
$$

Summing (8.8) over $k = 1, \ldots, B$, we obtain

$$\begin{aligned}
2\eta(1-2\eta\beta)\sum_{k=1}^{B}(\mathbb{E}\left[f(y_k) \mid x_t\right] - f(x^*)) &\leq \mathbb{E}\left[\|y_1 - x^*\|_2^2 \mid x_t\right] - \mathbb{E}\left[\|y_{B+1} - x^*\|_2^2 \mid x_t\right] \\
&\quad + 4\eta^2\beta B(f(x_t) - f(x^*)) \\
&\leq \|x_t - x^*\|_2^2 + 4\eta^2\beta B(f(x_t) - f(x^*)) \\
&\leq \left(\frac{2}{\alpha} + 4\eta^2\beta B\right)(f(x_t) - f(x^*)).
\end{aligned} \tag{8.9}$$

Dividing each side of (8.9) by B,

$$\begin{aligned}
2\eta(1-2\eta\beta)(\mathbb{E}\left[f\left(x_{t+1}\right) \mid x_t\right] - f(x^*)) &= 2\eta(1-2\eta\beta)(\mathbb{E}\left[f\left(\frac{1}{B}\sum_{k=1}^{B} y_k\right) \mid x_t\right] - f(x^*)) \\
&\leq 2\eta(1-2\eta\beta)\frac{1}{B}\sum_{k=1}^{B}(\mathbb{E}\left[f\left(y_k\right) \mid x_t\right] - f(x^*)) \\
&\leq \left(\frac{2}{\alpha B} + 4\eta^2\beta\right)(f(x_t) - f(x^*)).
\end{aligned} \tag{8.10}$$

Remember that

$$\eta = \frac{1}{6\beta}, \quad B = \frac{36\beta}{\alpha}.$$

Then it follows from (8.10) that

$$\begin{aligned}
\mathbb{E}\left[f\left(x_{t+1}\right) \mid x_t\right] - f(x^*)) &\leq \frac{1}{2\eta(1-2\eta\beta)}\left(\frac{2}{\alpha B} + 4\eta^2\beta\right)(f(x_t) - f(x^*)) \\
&= \frac{3\beta}{1-1/3}\left(\frac{1}{18\beta} + \frac{1}{9\beta}\right)(f(x_t) - f(x^*)) \\
&= \frac{3}{4}(f(x_t) - f(x^*)).
\end{aligned} \tag{8.11}$$

Applying the tower rule to (8.11),

$$\begin{aligned}
\mathbb{E}\left[f\left(x_{t+1}\right)\right] - f(x^*) &\leq \frac{3}{4}(\mathbb{E}\left[f(x_t)\right] - f(x^*)) \\
&\leq \left(\frac{3}{4}\right)^t (f(x_1) - f(x^*)),
\end{aligned} \tag{8.12}$$

as required. □

References

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 4

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 4

[SLRB17] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017. 4