23 Variance Reduction

Goal

Sampling, bias and variance, finite sum, variance reduction, SVRG, IG, IAG, SAG

Alert 23.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 23.2: Problem

In this lecture we continue our discussion of the most common problem in ML:

$$\min_{\mathbf{w}} \ \underbrace{\ell(\mathbf{w}) + r(\mathbf{w})}_{f(\mathbf{w})}, \quad \text{where} \quad \ell(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}),$$

where each ℓ_i and r are (closed) convex functions. Our main interest is the setting where n is extremely large, so that naively computing the (sub)gradient at each iteration is likely infeasible.

As mentioned in Example 22.3, we can apply stochastic gradient algorithms, where in each iteration t we randomly sample a minibatch $I_t = \lbag i_1, \ldots, i_m \rbag \subseteq \{1, \ldots, n\}$ of functions and update with the stochastic (sub)gradient:

$$\hat{\partial}\ell(\mathbf{w}_t) := \frac{1}{|I_t|}\sum_{i \in I_t} \partial\ell_i(\mathbf{w}_t) \approx \frac{1}{n}\sum_{i=1}^{n} \partial\ell_i(\mathbf{w}_t) =: \partial\ell(\mathbf{w}_t).$$

Remark 23.3: The bias and variance

We may think of each minibatch as a random set of size m (or in more fancy language, random counting measure or point process). We define its intensity

$$\mu_{i,t} = \mathbb{E} I_t(i),$$

where $I_t(i)$ is the random number of repetitions of ℓ_i in our minibatch I_t (of size m) at iteration t. Then,

$$\mathbb{E}\hat{\partial}\ell(\mathbf{w}_t) := \frac{1}{m}\mathbb{E}\left[\sum_{i \in I_t} \partial\ell_i(\mathbf{w}_t)\right] = \frac{1}{m}\sum_{i=1}^{n} \mu_{i,t}\partial\ell_i(\mathbf{w}_t).$$

Thus, as long as $\mu_{i,t} \equiv m/n$ we obtain an unbiased estimate of the (sub)gradient. Similarly, let

$$s_{i,j,t} = \mathbb{E} I_t(i) I_t(j) \leq \sqrt{s_{i,t} s_{j,t}}, \quad \text{where} \quad s_{i,t} := s_{i,i,t} = \mathbb{E} I_t^2(i).$$

Then, we also have

$$\begin{aligned}
\mathbb{E}\|\hat{\partial}\ell(\mathbf{w}_t)\|_2^2 &:= \frac{1}{m^2}\mathbb{E}\Big\|\sum_{i \in I_t} \partial\ell_i(\mathbf{w}_t)\Big\|_2^2 = \frac{1}{m^2}\sum_{i,j=1}^{n} s_{i,j,t}\,\langle \partial\ell_i(\mathbf{w}_t), \partial\ell_j(\mathbf{w}_t)\rangle \\
&\leq \frac{1}{m^2}\sum_{i,j} \sqrt{s_{i,t}s_{j,t}}\|\partial\ell_i(\mathbf{w}_t)\|_2 \cdot \|\partial\ell_j(\mathbf{w}_t)\|_2 = \frac{1}{m^2}\left(\sum_i \sqrt{s_{i,t}}\|\partial\ell_i(\mathbf{w}_t)\|_2\right)^2 \\
&\leq \frac{1}{m^2}\sum_i s_{i,t}\|\partial\ell_i(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

Exercise 23.4: Sampling w/o replacement

The following three sampling schemes are usually used in practice:

- Sampling with replacement. Verify that

  $$I_t \equiv \sum_{k=1}^{m} \delta_{Z_k},$$

  where Z_k's are i.i.d. uniformly random sample from $\{1, \ldots, n\}$, and δ_Z is the delta mass such that $\delta_Z(A) = 1$ if $Z \in A$ and 0 otherwise. It then follows that

  $$\mu_{i,t} = \mathbb{E}I_t(i) = m/n, \quad s_{i,t} = m/n + m(m-1)/n^2 \leq 2m/n, \quad \forall i \neq j,\ s_{i,j,t} = m(m-1)/n^2 \implies$$
  $$\mathbb{E}\hat{\partial}\ell(\mathbf{w}_t) = \partial\ell(\mathbf{w}_t), \quad \mathbb{E}\|\hat{\partial}\ell(\mathbf{w}_t)\|_2^2 \leq \frac{2}{m} \cdot \frac{1}{n}\sum_{i=1}^{n} \|\partial\ell_i(\mathbf{w}_t)\|_2^2.$$

  This is the most common and convenient scheme as we need only draw the m minibatch samples independently and identically. See Zhou et al. (2018) for some interesting extension.

- Sampling without replacement, in which case

  $$\mu_{i,t} = \binom{n-1}{m-1} / \binom{n}{m} = \frac{m}{n}, \quad s_{i,t} = \mu_{i,t} = \frac{m}{n}, \quad \forall i \neq j,\ s_{i,j,t} = \binom{n-2}{m-2} / \binom{n}{m} = \frac{m(m-1)}{n(n-1)} \leq \frac{m}{n}$$
  $$\mathbb{E}\hat{\partial}\ell(\mathbf{w}_t) = \partial\ell(\mathbf{w}_t), \quad \mathbb{E}\|\hat{\partial}\ell(\mathbf{w}_t)\|_2^2 \leq \frac{1}{m} \cdot \frac{1}{n}\sum_{i=1}^{n} \|\partial\ell_i(\mathbf{w}_t)\|_2^2.$$

  See Shamir (2016) for some interesting analysis.

- Randomly permuting the n functions followed by taking the n/m consecutive blocks as minibatches. This scheme empirically behaves similarly to sampling without replacement. See Gürbüzbalaban et al. (2019) for some interesting analysis.

Thus, we see that we can obtain unbiased estimate of the gradient while the size of the minibatch reduces the variance proportionally. However, inspecting Theorem 22.6 we see that reducing the variance helps improve the constant, but it does not seem to affect the $O(1/\sqrt{t})$ rate of convergence.

Zhou, P., X. Yuan, and J. Feng (2018). "New Insight into Hybrid Stochastic Gradient Descent: Beyond With-Replacement Sampling and Convexity". In: *Advances in Neural Information Processing Systems 31*.

Shamir, O. (2016). "Without-Replacement Sampling for Stochastic Gradient Methods". In: *Advances in Neural Information Processing Systems*, pp. 46–54.

Gürbüzbalaban, M., A. Ozdaglar, and P. A. Parrilo (2019). "Why random reshuffling beats stochastic gradient descent". *Mathematical Programming*.

Theorem 23.5: Faster rate under strong convexity

Under the same setting as in Remark 22.8, if f is L-Lipschitz continuous and σ-strongly convex (w.r.t. the norm $\|\cdot\|_2$), and the noise in (sub)gradient has variance bounded by ς^2, then with $\eta_t = \frac{1}{\sigma(t+1)}$ we have

$$\min_{0 \leq t \leq T-1} \mathbb{E}[f(\mathbf{w}_t) - f(\mathbf{w})] \leq \sum_{t=0}^{T-1} \frac{1}{T}\mathbb{E}[f(\mathbf{w}_t) - f(\mathbf{w})] \leq \frac{(\mathsf{L}^2 + \varsigma^2)\ln(T+1)}{2\sigma T}.$$

Proof: The proof is similar to that of Theorem 5.18. Conditioned on $\mathbf{w}_t$:

$$\mathbb{E}\,\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2\mathbb{E}\,\|\hat{\mathbf{w}}_t^*\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}, \mathbb{E}\hat{\mathbf{w}}_t^* \rangle$$
$$\text{[unbiasedness]}\ ☜ = \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2[\|\mathbb{E}\hat{\mathbf{w}}_t^*\|^2 + \mathrm{Var}(\hat{\mathbf{w}}_t^*)] - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}, \mathbf{w}_t^* \rangle$$

$$\begin{aligned} [\sigma\text{-strong convexity}] \ \text{☜} \ &\le \ (1-\sigma\eta_t)\,\|\mathbf{w}_t-\mathbf{w}\|_2^2+\eta_t^2[\|\mathbf{w}_t^*\|^2+\mathrm{Var}(\hat{\mathbf{w}}_t^*)]+2\eta_t(f(\mathbf{w})-f(\mathbf{w}_t)) \\ [\partial f \text{ is bounded by } \mathsf{L}] \ \text{☜} \ &\le \ \frac{t}{t+1}\,\|\mathbf{w}_t-\mathbf{w}\|_2^2+\eta_t^2(\mathsf{L}^2+\varsigma^2)+2\eta_t(f(\mathbf{w})-f(\mathbf{w}_t)). \end{aligned}$$

Telescoping we obtain

$$T\mathbb{E}\,\|\mathbf{w}_T-\mathbf{w}\|_2^2 \le \frac{\mathsf{L}^2+\varsigma^2}{\sigma^2}\sum_{t=0}^{T-1}\frac{1}{t+1}+\frac{2}{\sigma}\sum_{t=0}^{T-1}\mathbb{E}[f(\mathbf{w})-f(\mathbf{w}_t)].$$

Thus,

$$\min_{0\le t\le T-1}\mathbb{E}[f(\mathbf{w}_t)-f(\mathbf{w})] \le \sum_{t=0}^{T-1}\frac{1}{T}\mathbb{E}[f(\mathbf{w}_t)-f(\mathbf{w})] \le \frac{(\mathsf{L}^2+\varsigma^2)\sum_{t=0}^{T-1}\frac{1}{t+1}}{2\sigma T} \le \frac{(\mathsf{L}^2+\varsigma^2)\ln(T+1)}{2\sigma T},$$

as claimed. ■

If we define $\bar{\mathbf{w}}_T=\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}_t$, then obviously

$$\mathbb{E}[f(\bar{\mathbf{w}}_T)-f(\mathbf{w})] \le \sum_{t=0}^{T-1}\frac{1}{T}\mathbb{E}[f(\mathbf{w}_t)-f(\mathbf{w})] \le \frac{(\mathsf{L}^2+\varsigma^2)\ln(T+1)}{2\sigma T}.$$

Exercise 23.6: Stochastic GDA under strong convexity

Extend Theorem 23.5 to the stochastic gradient descent ascent algorithm for any monotone VI.

Alert 23.7: Grave danger of wrong parameter

What if we do not know σ and unfortunately overestimate it? The following example from Nemirovski et al. (2009) is quite illuminating.

Consider $f(w)=w^2/10$ (so that $\sigma=1/5$) and $C=[-1,1]$. Suppose we set $\eta_t=1/(t+1)$. Then

$$w_{t+1}=w_t-\tfrac{1}{t+1}\tfrac{1}{5}w_t=\left(1-\tfrac{1}{5(t+1)}\right)w_t.$$

Thus, with $w_0=1$ we have

$$w_t=\prod_{s=1}^{t}(1-\tfrac{1}{5s})=\exp\left\{-\sum_{s=1}^{t}\ln\left(1+\tfrac{1}{5s-1}\right)\right\}>\exp\left\{-\sum_{s=1}^{t}\tfrac{1}{5s-1}\right\}>0.8(t+1)^{-1/5},$$

which is even slower than the $O(1/\sqrt{t})$ rate we obtained in Remark 22.8 without strong convexity!

Of course, if we guessed the strong convexity parameter σ correctly and used $\eta_t=5/(t+1)$, the algorithm would converge to the minimizer 0 in a single iteration! Unfortunately, it is not easy to line search σ, especially in the presence of stochastic noise.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609.

Algorithm 23.8: Stochastic variance reduced gradient (SVRG, Johnson and Zhang 2013)

```
Algorithm: Stochastic variance reduced proximal gradient
Input: $\mathbf{w}_0 \in \mathrm{dom}\, f$
for $k = 0, 1, 2, \ldots$ do
    $\mathbf{g}_k \leftarrow \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(\mathbf{w}_k)$          // compute the full gradient at epoch $k$
    $\mathbf{w}_{k,0} \leftarrow \mathbf{w}_k$
    for $t = 0, \ldots, m-1$ do
        randomly draw $i_t = i$ with probability $p_i$
        $\mathbf{g}_{k,t} \leftarrow \mathbf{g}_k - \frac{1}{np_{i_t}}\nabla \ell_{i_t}(\mathbf{w}_k) + \frac{1}{np_{i_t}}\nabla \ell_{i_t}(\mathbf{w}_{k,t})$   // update gradient in amortized fashion
        $\mathbf{w}_{k,t+1} \leftarrow \mathrm{P}_r^{\eta_k}(\mathbf{w}_{k,t} - \eta_k \mathbf{g}_{k,t})$          // stochastic proximal gradient
    $\mathbf{w}_{k+1} \leftarrow \frac{1}{m}\sum_{t=1}^m \mathbf{w}_{k,t}$          // in practice, can also do $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_{k,m}$
```

The above algorithm, with $r \equiv 0$, is due to Johnson and Zhang (2013) and later extended by Xiao and Zhang (2014) to any convex r whose proximal map can be easily computed. The main idea is to amortize the computation of full gradient. Compared to vanilla stochastic gradient, on average SVRG requires computing 2 gradients per step (3, if we choose to recompute each $\nabla \ell_{i_t}(\mathbf{w}_k)$ instead of storing them).

Let us note that the stochastic gradient used in SVRG is still unbiased:

$$\mathbb{E}\mathbf{g}_{k,t} = \mathbf{g}_k + \sum_{i=1}^n p_i \cdot \tfrac{1}{np_i}\left[-\nabla \ell_i(\mathbf{w}_k) + \nabla \ell_i(\mathbf{w}_{k,t})\right] = \mathbf{g}_k - \mathbf{g}_k + \tfrac{1}{n}\sum_{i=1}^n \nabla \ell_i(\mathbf{w}_{k,t}) = \nabla \ell(\mathbf{w}_{k,t}).$$

Moreover, if $\mathbf{w}_k \approx \mathbf{w}_{k,t}$, e.g. when the algorithm is close to convergence, the variance of $\mathbf{g}_{k,t}$ will be small (since the random fluctuations cancel each other). Indeed, let $L = \max_i L_i/(np_i)$ where ℓ_i is L_i-smooth. Then,

$$\begin{aligned}
\mathbb{E}\|\mathbf{g}_{k,t} - \mathbb{E}\mathbf{g}_{k,t}\|_2^2 &\le \mathbb{E}\|\tfrac{1}{np_{i_t}}[\nabla \ell_{i_t}(\mathbf{w}_{k,t}) - \nabla \ell_{i_t}(\mathbf{w}_k)]\|_2^2 = \sum_{i=1}^n \tfrac{1}{n^2 p_i}\|\nabla \ell_i(\mathbf{w}_{k,t}) - \nabla \ell_i(\mathbf{w}_k)\|_2^2 \\
[(a+b)^2 \le 2(a^2+b^2)] \text{ ☜ } &\le 4L \sum_{i=1}^n \tfrac{1}{2nL_i}[\|\nabla \ell_i(\mathbf{w}_{k,t}) - \nabla \ell_i(\mathbf{w}_\star)\|_2^2 + \|\nabla \ell_i(\mathbf{w}_k) - \nabla \ell_i(\mathbf{w}_\star)\|_2^2] \\
[\text{Alert 6.25}] \text{ ☜ } &\le 4L \sum_{i=1}^n \tfrac{1}{n}[\mathsf{D}_{\ell_i}(\mathbf{w}_{k,t}, \mathbf{w}_\star) + \mathsf{D}_{\ell_i}(\mathbf{w}_k, \mathbf{w}_\star)] = 4L[\mathsf{D}_{\ell}(\mathbf{w}_{k,t}, \mathbf{w}_\star) + \mathsf{D}_{\ell}(\mathbf{w}_k, \mathbf{w}_\star)] \\
&\le 4L[f(\mathbf{w}_{k,t}) - f(\mathbf{w}_\star) + f(\mathbf{w}_k) - f(\mathbf{w}_\star)],
\end{aligned}$$

where we applied Proposition 4.20 to $\mathbf{w}_\star \in \mathrm{argmin}\, f$ in the last line.

Johnson, R. and T. Zhang (2013). "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems*.

Xiao, L. and T. Zhang (2014). "A Proximal Stochastic Gradient Method with Progressive Variance Reduction". *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075.

Lemma 23.9: Inexact proximal gradient

Let ℓ be L-smooth convex and r be convex. For any $\eta \in (0, 1/L]$ define

$$\mathbf{w}^+ = \mathrm{P}_r^\eta\big(\mathbf{w} - \eta(\nabla \ell(\mathbf{w}) + \boldsymbol{\varepsilon})\big) = \operatorname*{argmin}_{\mathbf{z}} \ \langle \mathbf{z}, \nabla \ell(\mathbf{w}) + \boldsymbol{\varepsilon}\rangle + \tfrac{1}{2\eta}\|\mathbf{z} - \mathbf{w}\|_2^2 + r(\mathbf{z}).$$

Then, for any $\mathbf{z}$ we have

$$f(\mathbf{z}) \ge f(\mathbf{w}^+) + \tfrac{1}{\eta}\big\langle \mathbf{w} - \mathbf{w}^+, \mathbf{z} - \mathbf{w}\big\rangle + \tfrac{1}{2\eta}\|\mathbf{w} - \mathbf{w}^+\|_2^2 - \big\langle \mathbf{z} - \mathbf{w}^+, \boldsymbol{\varepsilon}\big\rangle \tag{23.1}$$

Proof: We apply Proposition 4.20 to $\mathbf{w}^+$:

$$\langle \mathbf{z}, \nabla\ell(\mathbf{w}) + \boldsymbol{\varepsilon}\rangle + \tfrac{1}{2\eta}\|\mathbf{z}-\mathbf{w}\|_2^2 + r(\mathbf{z}) \geq \left\langle \mathbf{w}^+, \nabla\ell(\mathbf{w}) + \boldsymbol{\varepsilon}\right\rangle + \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{w}^+\|_2^2 + r(\mathbf{w}^+) + \tfrac{1}{2\eta}\|\mathbf{z}-\mathbf{w}^+\|_2^2.$$

Adding the following inequalities from L-smoothness and convexity:

$$\ell(\mathbf{w}) + \left\langle \mathbf{w}^+ - \mathbf{w}, \nabla\ell(\mathbf{w})\right\rangle + \tfrac{1}{2\eta}\|\mathbf{w}-\mathbf{w}^+\|_2^2 \geq \ell(\mathbf{w}^+)$$
$$\ell(\mathbf{z}) \geq \ell(\mathbf{w}) + \langle \mathbf{z}-\mathbf{w}, \nabla\ell(\mathbf{w})\rangle,$$

and rearranging and simplifying leads to (23.1). ■

When ℓ or r are strongly convex, we can sharpen the bound (23.1), although this is not needed below. When $\boldsymbol{\varepsilon} = \mathbf{0}$, i.e. the gradient is exact, the resulting bound has been obtained and used before (e.g. with $\mathbf{z} = \mathbf{w}$).

Theorem 23.10: Linear convergence of SVRG (Xiao and Zhang 2014)

Let ℓ_i be L_i smooth convex and f be σ-strongly convex. Let $L = \max_i L_i/(np_i)$ and $\eta_k \equiv \eta \in (0, 1/(4L))$. Then, for the epoch updates:

$$\mathbb{E}[f(\mathbf{w}_{k+1}) - f_\star] \leq c\mathbb{E}[f(\mathbf{w}_k) - f_\star], \quad \textit{where} \quad c = \frac{1/(\eta\sigma) + 4\eta L(m+1)}{(1-4\eta L)m}.$$

Note that $c < 1$ if m is sufficiently large.

Proof: Let $\mathbf{w}_\star = \operatorname{argmin} f$ and we bound the progress of SVRG in the inner-loop as usual:

$$\|\mathbf{w}_{k,t+1} - \mathbf{w}_\star\|_2^2 = \|\mathbf{w}_{k,t} - \mathbf{w}_\star\|_2^2 + 2\left\langle \mathbf{w}_{k,t} - \mathbf{w}_{k,t+1}, \mathbf{w}_\star - \mathbf{w}_{k,t}\right\rangle + \|\mathbf{w}_{k,t+1} - \mathbf{w}_{k,t}\|_2^2$$
$$[\text{Lemma 23.9}] ☜ \leq \|\mathbf{w}_{k,t} - \mathbf{w}_\star\|_2^2 - 2\eta_k[f(\mathbf{w}_{k,t+1}) - f(\mathbf{w}_\star)] + 2\eta_k\left\langle \mathbf{w}_{k,t+1} - \mathbf{w}_\star, \nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\right\rangle$$

Let $\tilde{\mathbf{w}}_{k,t+1} := \mathrm{P}_r^{\eta_k}(\mathbf{w}_{k,t} - \eta_k\nabla\ell(\mathbf{w}_{k,t}))$ we can continue bounding

$$\begin{aligned}\langle \mathbf{w}_{k,t+1} - \mathbf{w}_\star, \nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\rangle &\leq \langle \mathbf{w}_{k,t+1} - \tilde{\mathbf{w}}_{k,t+1}, \nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\rangle + \langle \tilde{\mathbf{w}}_{k,t+1} - \mathbf{w}_\star, \nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\rangle \\ &\leq \|\mathbf{w}_{k,t+1} - \tilde{\mathbf{w}}_{k,t+1}\|_2 \cdot \|\nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\|_2 + \langle \tilde{\mathbf{w}}_{k,t+1} - \mathbf{w}_\star, \nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\rangle \\ &\leq \eta_k\|\nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\|_2^2 + \langle \tilde{\mathbf{w}}_{k,t+1} - \mathbf{w}_\star, \nabla\ell(\mathbf{w}_{k,t}) - \mathbf{g}_{k,t}\rangle.\end{aligned}$$

Taking expectations and noting that $\tilde{\mathbf{w}}_{k,t+1}$ does not depend on i_t, we apply the unbiasedness and variance bound in Line 8:

$$\mathbb{E}\|\mathbf{w}_{k,t+1} - \mathbf{w}_\star\|_2^2 \leq \mathbb{E}\|\mathbf{w}_{k,t} - \mathbf{w}_\star\|_2^2 - 2\eta_k\mathbb{E}[f(\mathbf{w}_{k,t+1}) - f_\star] + 8L\eta_k^2\mathbb{E}[f(\mathbf{w}_{k,t}) - f_\star + f(\mathbf{w}_k) - f_\star]$$

Summing over t from 0 to $m-1$ and noting that $\mathbf{w}_{k,0} = \mathbf{w}_k$:

$$\mathbb{E}\|\mathbf{w}_{k,m} - \mathbf{w}_\star\|_2^2 \leq \mathbb{E}\|\mathbf{w}_k - \mathbf{w}_\star\|_2^2 - 2\eta_k\sum_{t=0}^{m-1}\mathbb{E}[f(\mathbf{w}_{k,t+1}) - f_\star] + 8L\eta_k^2\sum_{t=0}^{m-1}\mathbb{E}[f(\mathbf{w}_{k,t}) - f_\star + f(\mathbf{w}_k) - f_\star].$$

Rearranging and using the definition of $\mathbf{w}_{k+1}$:

$$\begin{aligned}2\eta_k(1-4\eta_k L)m\mathbb{E}[f(\mathbf{w}_{k+1}) - f_\star] &\leq \mathbb{E}\|\mathbf{w}_k - \mathbf{w}_\star\|_2^2 + 8\eta_k^2 L(m+1)\mathbb{E}[f(\mathbf{w}_k) - f_\star] \\ &\leq \big(2/\sigma + 8\eta_k^2 L(m+1)\big)\mathbb{E}[f(\mathbf{w}_k) - f_\star].\end{aligned}$$

Dividing the constant we obtain the formula for c and the proof is complete. ■

If we let $\eta = 1/\sqrt{4\sigma L(m+1)}$ with $m + 1 > 4\kappa$, where $\kappa := L/\sigma \geq 1$ is the condition number, then

$$c = \frac{4\sqrt{\kappa}}{\sqrt{m+1} - 2\sqrt{k}} \cdot \frac{m+1}{m},$$

leading to the expected overall complexity $O\big((n+\kappa)\log\frac{1}{\epsilon}\big)$ for an ϵ-approximation minimizer.

Xiao, L. and T. Zhang (2014). "A Proximal Stochastic Gradient Method with Progressive Variance Reduction". *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075.

Exercise 23.11: Non-uniform vs. uniform sampling

To minimize $L = \max_i L_i/(np_i)$, we solve

$$\min_{\mathbf{p}\in\Delta} \max_i L_i/p_i.$$

Prove that the optimal $p_i \propto L_i$, leading to $L = \frac{1}{n}\sum_i L_i$, which makes intuitive sense: the more "curvy" (i.e. a large L_i) a component function is, the more attention we pay to it.

In contrast, if we set $p_i \equiv 1/n$, then $L = \max_i L_i$, which is strictly larger.

When L_i's are not available or expensive to estimate, we may use the successive difference of the gradients to approximate it; recall the line search procedure of Khobotov in Line 6.

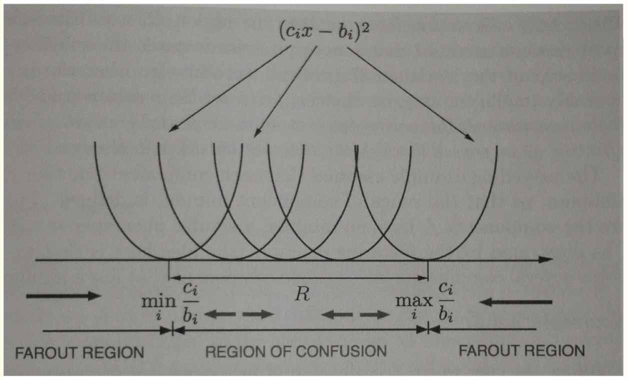
Algorithm 23.12: Incremental gradient (IG, e.g. Bertsekas 2011)

```
Algorithm: Incremental gradient (IG)
Input: $\mathbf{w} \in \operatorname{dom} f$
for $k = 0, 1, 2, \ldots$ do
    for $t = 0, \ldots, m-1$ do
        choose $i_t$                                                    // cyclic or random
        $\mathbf{w} \leftarrow \mathrm{P}^{\eta_k}_{r_{i_t}}(\mathbf{w} - \eta_k \nabla \ell_{i_t}(\mathbf{w}))$          // proximal gradient on component $\ell_{i_t} + r_{i_t}$
```

For simplicity, let us assume $r = \sum_i r_i \equiv 0$ and we choose the cyclic rule (hence $m = n$). Then, we may write the inner loop compactly as:

$$\begin{aligned}
\mathbf{w}_{k+n} &= \mathbf{w}_k - \eta_k \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}_{k+i-1}), \quad \text{where} \quad \mathbf{w}_{k+i} = \mathbf{w}_{k+i-1} - \eta_k \tfrac{1}{n}\nabla \ell_i(\mathbf{w}_{k+i-1}) \\
&= \mathbf{w}_k - \eta_k \nabla \ell(\mathbf{w}_k) + \eta_k \underbrace{\frac{1}{n}\sum_{i=1}^{n}[\nabla \ell_i(\mathbf{w}_k) - \nabla \ell_i(\mathbf{w}_{k+i-1})]}_{\boldsymbol{\varepsilon}_k}
\end{aligned}$$

If ℓ_i's are L-smooth and $\eta_k \to 0$, then it is possible for the gradient error to diminish, see e.g. Bertsekas (2011) and Lan and Zhou (2018).



Fix: diminishing stepsize, or live with small error

Bertsekas, D. P. (2011). "Incremental proximal methods for large scale convex optimization". *Mathematical Programming*, vol. 129, pp. 163–195.

Lan, G. and Y. Zhou (2018). "Random Gradient Extrapolation for Distributed and Stochastic Optimization". *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 2753–2782.

Algorithm 23.13: Incremental/Stochastic averaged gradient (I/SAG, e.g. Blatt et al. 2007)

```
Algorithm: Incremental/stochastic averaged gradient (I/SAG)
Input: $\mathbf{w}_0 \in \operatorname{dom} f$, $G \in \mathbb{R}^{d\times n}$
$\mathbf{g}_{-1} \leftarrow \frac{1}{n} G\mathbf{1}$                          // $G$ stores most recent gradient for each $\ell_i$
for $t = 0, 1, 2, \ldots$ do
    choose $i_t$                                                  // cyclic or random
    $\mathbf{g}_t \leftarrow \mathbf{g}_{t-1} - \frac{1}{n} G_{:,i_t} + \frac{1}{n}\nabla \ell_{i_t}(\mathbf{w}_t)$            // replace old with new
    $G_{:,i_t} \leftarrow \nabla \ell_{i_t}(\mathbf{w}_t)$
    $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_r^{\eta_t}(\mathbf{w}_t - \eta_t \mathbf{g}_t)$                  // inexact proximal gradient
```

When we update the component functions ℓ_i sequentially, the update may be written more compactly as:

$$\mathbf{w}_{t+1} \leftarrow \mathrm{P}_r^{\eta_t}\left(\mathbf{w}_t - \eta_t \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_{i_{t-n+i}}(\mathbf{w}_{t-n+i})\right)$$

The sequential version was analyzed in Gürbüzbalaban et al. (2017), Mokhtari et al. (2018), Vanli et al. (2018), and Gürbüzbalaban et al. (2019) while the randmized version in Schmidt et al. (2017), achieving similar rates of convergence as SVRG (see Theorem 23.10).

Blatt, D., A. O. Hero, and H. Gauchman (2007). “A Convergent Incremental Gradient Method with a Constant Step Size”. *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 29–51.

Gürbüzbalaban, M., A. Ozdaglar, and P. A. Parrilo (2017). “On the Convergence Rate of Incremental Aggregated Gradient Algorithms”. *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 1035–1048.

Mokhtari, A., M. Gürbüzbalaban, and A. Ribeiro (2018). “Surpassing Gradient Descent Provably: A Cyclic Incremental Method with Linear Convergence Rate”. *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1420–1447.

Vanli, N. D., M. Gürbüzbalaban, and A. Ozdaglar (2018). “Global Convergence Rate of Proximal Incremental Aggregated Gradient Methods”. *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1282–1300.

Gürbüzbalaban, M., A. Ozdaglar, and P. A. Parrilo (2019). “Why random reshuffling beats stochastic gradient descent”. *Mathematical Programming*.

Schmidt, M., N. L. Roux, and F. Bach (2017). “Minimizing finite sums with the stochastic average gradient”. *Mathematical Programming*, vol. 162, pp. 83–112.