Today:

Revisiting an older method for minimization that regained popularity.

Randomized Coordinate Descent:

$$w_{k+1} = w_k - \gamma [\nabla f(w)]_{i_k}$$

where i_k is picked:

- uniformly at random
- random with importance weights
- greedily.

Q: How does RCD compare with GD?

We will assume the following coordinate-wise smoothness:

<u>Property</u>: A function $f(w)$ is β_i- coordinate-wise smooth if:

$$\forall a, x, i \quad f(x + a e_i) \leq f(x) + a[\nabla f(x)]_i + \frac{\beta_i a^2}{2}$$

Let $\beta = \max_i \beta_i$.

Further let us assume that $f(w)$ is μ- PL:

$$\forall w, \quad \|\nabla f(w)\|^2 \geq \mu/2 \,(f(w) - f^*)$$

Then, we can show the following guarantees for RCD with uniform sampling

Thm: If f is β-coordinate-wise smooth and μ-PL then RCD with stepsize $1/L$:

$$w_{k+1} = w_k - \frac{1}{\beta}\left[\nabla f(w_k)\right]_{i_k}$$

$$i_k \sim \mathrm{unif}(1, \dots, n)$$

obtains the following rate:

$$\mathbb{E}\, f(w_k) - f^* \le \left(1 - \frac{\mu}{d\beta}\right)^k \left(f(x_0) - f^*\right)$$

Proof:

By plugging

$$w_{k+1} = w_k - \frac{1}{\beta}\left[\nabla f(w_k)\right]_{i_k}$$

In the coordinate-wise smoothness property, we get

$$f(\omega_{k+1}) \leq f(\omega_k) - \frac{1}{2\beta}\left|[\nabla f(\omega_k)]_{i_k}\right|^2$$

By taking expectation (w.r.t i_k), we have:

$$\begin{aligned} E f(\omega_{k+1}) &\leq f(\omega_k) - \frac{1}{2\beta} E\left|[\nabla f(\omega_k)]_{i_k}\right|^2 \\ &= f(\omega_k) - \frac{1}{2\beta}\sum_{i=1}^{d}\frac{1}{d}\left|[\nabla f(\omega_k)]_i\right|^2 \\ &= f(\omega_k) - \frac{1}{2\beta d}\|\nabla f(\omega_k)\|^2 \end{aligned}$$

We will now apply the PL condition to get:

$$E f(\omega_{k+1}) \leq f(\omega_k) - \frac{\mu}{2\beta d}\left(f(\omega_k) - f^*\right)$$

$\Rightarrow$

$$\mathbb{E} f(\omega_{k+1}) - f^* \leq f(\omega_k) - f^* - \frac{\mu}{2\beta d}\left(f(\omega_k) - f^*\right)$$

$$= \left(1 - \frac{\mu}{\beta d}\right) f(\omega_k) - f^*$$

Applying expectations and recursively expanding yields the result.

$\boxtimes$

Unfortunately, as it stands RCD with uniform sampling is not clearly better than CD which achieves a rate

$$f(\omega_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_0) - f^*\right)$$

which requires $O(d)$ less iterations than RCD for the same accuracy ε.

To improve RCD we can employ <u>importance sampling</u>:

$$\Pr(i_k = i) = \frac{\beta_i}{\sum_{j=1}^{d} \beta_j}$$

That is, each coordinate is sampled proportionally to its "effect" on $f(w)$.

For this weighted sampling we get:

$$w_{k+1} = w_k - \frac{1}{\beta_i} [\nabla f(w_k)]_{i_k}$$

<u>and</u>

$$E f(w_{k+1}) \leq f(w_k) - \frac{1}{\beta_i} E \left| [\nabla f(w_k)] \right|_{i_k}$$

$$= f(w_k) - \frac{1}{\beta_i} \frac{\beta_i}{\sum \beta_i} \frac{1}{d} \|\nabla f(w_k)\|^2$$

$$\leq f(w_k) - \frac{1}{\bar{\beta} \cdot d} \|\nabla f(w_k)\|^2$$

The above imply the following Theorem:

<u>Thm:</u> RCD with importance sampling according to

$$\Pr(i_k = i) = \frac{\beta_i}{\sum_j \beta_j}$$

yields the following rate:

$$\mathbb{E} f(w_k) - f^* \leq \left(1 - \frac{\mu}{d\bar{\beta}}\right)^k (f(w_0) - f^*)$$

<u>Remark:</u> $\bar{\beta}$ can be significantly smaller than β.

Example:

If we are aiming for ε-accuracy then:

$$T_\varepsilon^{\text{importance}} = O\left(\frac{d \sum_i \beta_i}{n \cdot \mu} \log\left(\frac{f_0 - f^*}{\varepsilon}\right)\right)$$

$$T_\varepsilon^{\text{uniform}} = O\left(\frac{d\, \beta_{max}}{\mu} \log(f_0 - f^*/\varepsilon)\right)$$

Hence importance sampling can be extremely helpful.

Main Question:

- How easy are β_i to compute?
- Similar sampling for SGD?