Today:

Revisiting an older method for minimization that regained popularity.

Randomized Coordinate Descent:

\[ w_{k+1} = w_k - \gamma \frac{\partial f(w)}{\partial w} i_k \]

where \( i_k \) is picked:

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

Q: How does RCD compare with CD?
We will assume the following coordinate-wise smoothness:

Property: A function \( f(x) \) is \( B_i \)-coordinate-wise smooth if:

\[ f(x + a e_i) \leq f(x) + a \nabla f(x)_i + \frac{B_i a^2}{2} \]

Let \( B = \max_i B_i \).

Further let us assume that \( f(w) \) is \( \mu \)-PL:

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

Then, we can show the following guarantees for RCD with uniform sampling.
Thus: If \( f \) is \( B \)-coordinate-wise smooth and \( \ell_1 \)-\( P \) then RCD with stepsize \( \frac{1}{L} \):

\[
W_{k+1} = W_k - \frac{1}{B} \left[ \nabla f(W_k) J_i k \right]
\]

\( i \in \text{unif}(1, \ldots, n) \)

obtains the following rate:

\[
\mathbb{E} f(W_k) - f^* \leq \left( 1 - \frac{1}{d \beta} \right)^k (f(x_0) - f^*)
\]

**Proof:**

By plugging

\[
W_{k+1} = W_k - \frac{1}{B} \left[ \nabla f(W_k) J_i k \right]
\]

In the coordinate-wise smoothness property, we get
\[ f(w_{k+1}) \leq f(w_k) - \frac{1}{2\beta} \left| \nabla f(w_k) \right|_{i_k}^2 \]

By taking expectation, we have:

\[ \mathbb{E} f(w_{k+1}) \leq f(w_k) - \frac{1}{2\beta} \mathbb{E} \left| \nabla f(w_k) \right|_{i_k}^2 \]

\[ = f(w_k) - \frac{1}{2\beta} \frac{1}{d} \sum_{i=1}^{d} \left| \nabla f(w_k) \right|_{i_k}^2 \]

\[ = f(w_k) - \frac{1}{2\beta d} \left\| \nabla f(w_k) \right\|^2 \]

We will now apply the PL condition to get:

\[ \mathbb{E} f(w_{k+1}) \leq f(w_k) - \frac{1}{2\beta d} \left( f(w_k) - f^* \right) \]
\[ Ef(\omega_{k+1})f^* \leq f(\omega_k) - f^* \frac{1}{2\beta_1} (f(\omega_k) - f^*) \]

\[ = (1 - \frac{1}{\beta_1}) f(\omega_k) - f^* \]

Applying expectations and recursively expanding yields the result.

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

\[ f(\omega_k) - f^* \leq (1 - \frac{1}{2}) \cdot \frac{3}{4} (f(x_0) - f^*) \]

which requires \( O(d) \) less iterations than RCD for the same accuracy \( \epsilon \).
To improve RCD we can employ importance sampling:

\[
Pr(\mathbf{i}k = i) = \frac{B_i}{\sum_{j=1}^{d} B_j}
\]

That is, each coordinate is sampled proportionally to its "effect" on \( f(\mathbf{w}) \).

For this weighted sampling we get:

\[
\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{1}{B_i} \left[ \nabla f(\mathbf{w}_k) \right]_{ik}
\]

and

\[
E f(\mathbf{w}_{k+1}) \leq f(\mathbf{w}_k) - \frac{1}{B_i} E \left[ \nabla f(\mathbf{w}_k) \right]_{ik}
\]

\[
= f(\mathbf{w}_k) - \frac{1}{B_i} \frac{B_i}{\sum B_i} \frac{1}{||\nabla f(\mathbf{w}_k)||^2}
\]
\[ \leq f(w_k) - \frac{1}{\bar{B}.d} \|\nabla f(w_k)\|^2 \]

The above imply the following theorem:

Thus, RCD with importance sampling according to

\[ \Pr(i_k = i) = \frac{B_i}{\sum_j B_j} \]

yields the following rate:

\[ E[f(w_k)] - f^* \leq (1 - \frac{h}{d\bar{B}})^k (f(w_0) - f^*) \]

Remark: \( \bar{B} \) can be significantly smaller than \( B \).
Example:

If we are aiming for $\varepsilon$-accuracy then:

$$T_{\text{importance}} = O( \frac{\delta \sum_i B_i}{n^{-1} n} \log \left( \frac{t_0 - t_F}{\varepsilon} \right))$$

$$T_{\text{uniform}} = O( \frac{d B_{\text{max}}}{\mu} \log \left( \frac{t_0 - t_F}{\varepsilon} \right))$$

Hence importance sampling can be extremely helpful.

Main Question:

- How easy are $B_i$ to compute?
- Similar sampling for SGD?