Chapter 5 Coordinate Descent

- Key Idea Update only one coordinate
- Results: Worse case d times more of iteration, under suitable condition results may improve.

Polyak-Łojasiewicz inequality

- *Goal* only to prove function values can converge to optimal, no care of complexity.
- Definition 5.1 Let $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ has a global minimum $\mathbf{x}^\star$. We say that f satisfies the *Polyak-Łojasiewicz inequality* (PL inequality) if $\exists \mu > 0$ s.t. $\forall \mathbf{x} \in \mathbb{R}^d$, $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu\left(f(\mathbf{x}) - f\left(\mathbf{x}^\star\right)\right)$.
- Lemma 5.2 (Strong Convexity -> PL ineq) The proof is in *Lemma E* of lecture 03.
- PL ineq is strictly weaker than strong convexity.
  - E.g. $f\left(x_1, x_2\right) = x_1^2$ is not strongly convex, but satisfies PL ineq.
  - Example of non-convex function that satisfies PL ineq: $f(x) = x^2 + \operatorname{Sigmoid}(\frac{x}{20}) \times 5x^2$.
- Theorem 5.3 (Exponential decay) Let $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ has a global minimum $\mathbf{x}^\star$. Suppose that f is L-smooth satiefies μ-PL ineq. Choosing stepsize $\gamma = L^{-1}$, then gradient descent starting with arbitrary $\mathbf{x}_0$ satiefies $f\left(\mathbf{x}_T\right) - f\left(\mathbf{x}^\star\right) \leq \left(1 - \frac{\mu}{L}\right)^T\left(f\left(\mathbf{x}_0\right) - f\left(\mathbf{x}^\star\right)\right)$
  - Proof
    - By sufficient descent $f\left(\mathbf{x}_{t+1}\right) \leq f\left(\mathbf{x}_t\right) - \frac{1}{2L}\left\|\nabla f\left(\mathbf{x}_t\right)\right\|^2$,
    - by PL ineq RHS $\leq f\left(\mathbf{x}_t\right) - \frac{\mu}{L}\left(f\left(\mathbf{x}_t\right) - f\left(\mathbf{x}^\star\right)\right)$, then we get what we want.

Coordinate Smoothness

- Definition 5.4 Let $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ and $\mathcal{L} = \left(L_1, L_2, \ldots, L_d\right) \in \mathbb{R}^d$. Function f is called *coordinate-wise smooth (with parameter $\mathcal{L}$)* if $\forall i \in [d]$, $\forall \mathbf{x} \in \mathbb{R}^d, \lambda \in \mathbb{R}, f\left(\mathbf{x} + \lambda \mathbf{e}_i\right) \leq f(\mathbf{x}) + \lambda \nabla_i f(\mathbf{x}) + \frac{L_i}{2}\lambda^2$
  - If $\forall i, L_i = L$, f is said to be coordinate-wise smooth with parameter L.
- Coordinate-wise smoothness is more fine grined.
  - $f\left(x_1, x_2\right) = x_1^2 + 10x_2^2$ is $(2, 20)$-coordinate-smooth, but only 20-smooth.
  - $f\left(x_1, x_2\right) = x_1^2 + x_2^2 + Mx_1x_2$ is $(2, 2)$-coordinate-smooth, but only $(M+2)$-smooth.
- Algorithm
  - (i) choose an active coordinate $i \in [d]$
  - (ii) $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f\left(\mathbf{x}_t\right)\mathbf{e}_i$
- Lemma 5.5 (Sufficient Descent) A stepsize of $\gamma_i = L_i^{-1}$ gives $f\left(\mathbf{x}_{t+1}\right) \leq f\left(\mathbf{x}_t\right) - \frac{1}{2L_i}\left|\nabla_i f\left(\mathbf{x}_t\right)\right|^2$.

Randomized CD

- Algorithm Change (i) sample $i \in [d]$ uniformly at random.
- Theorem 5.6 Let $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ has a global minimum $\mathbf{x}^\star$. Suppose that f is coordinate-wise L-smooth and satisfies the μ-PL inequality. Choosing stepsize $\gamma_i = L^{-1}$, then randomized coordinate descent with arbitrary start satisfies $\mathbb{E}\left[f\left(\mathbf{x}_T\right) - f\left(\mathbf{x}^\star\right)\right] \leq \left(1 - \frac{\mu}{dL}\right)^T\left(f\left(\mathbf{x}_0\right) - f\left(\mathbf{x}^\star\right)\right)$.
  - Proof
    - Take expectation over sufficient descent with each coordiate of probability $1/d$, $\mathbb{E}\left[f\left(\mathbf{x}_{t+1}\right) \mid \mathbf{x}_t\right] \leq f\left(\mathbf{x}_t\right) - \frac{1}{2L}\sum_{i=1}^d \frac{1}{d}\left|\nabla_i f\left(\mathbf{x}_t\right)\right|^2 = f\left(\mathbf{x}_t\right) - \frac{1}{2dL}\left\|\nabla f\left(\mathbf{x}_t\right)\right\|^2$
    - With PL ineq, RHS $\leq f\left(\mathbf{x}_t\right) - \frac{\mu}{dL}\left(f\left(\mathbf{x}_t\right) - f\left(\mathbf{x}^\star\right)\right)$
    - Take expectation over condition $\mathbb{E}[\cdot \mid \mathbf{x}_t]$, we arrive at our conclusion.
  - Comment $\left(1 - \frac{\mu}{L}\right) \approx \left(1 - \frac{\mu}{dL}\right)^d$, this is nearly the same as vanilla GD, need improvement.

Importance Sampling

- Algorithm Change (i) when L_i not the same, sample $i \in [d]$ with probability $\frac{L_i}{\sum_{j=1}^d L_j}$
- Theorem 5.7 Let $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ has a global minimum $\mathbf{x}^\star$. Suppose that f is coordinate-wise $\mathcal{L}$-smooth and satisfies the μ-PL inequality. Let $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$ be the average of all coordinate-wise smoothness constants. Then coordinate descent with importance sampling with arbitrary start satisfies $\mathbb{E}\left[f\left(\mathbf{x}_T\right) - f\left(\mathbf{x}^\star\right)\right] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^T\left(f\left(\mathbf{x}_0\right) - f\left(\mathbf{x}^\star\right)\right)$

- Proof
  - $\mathbb{E}\left[f\left(\mathbf{x}_{t+1}\right)|\mathbf{x}_t\right] \leq f\left(\mathbf{x}_t\right)-\sum_{i=1}^d \frac{1}{2 \bar{L}} \frac{L_i}{\sum_{j=1}^d L_j}\left|\nabla_i f\left(\mathbf{x}_t\right)\right|^2=f\left(\mathbf{x}_t\right)-\frac{1}{2 d \bar{L}}\left\|\nabla f\left(\mathbf{x}_t\right)\right\|^2$
- Comment $\bar{L}$ can be much smaller than $L=\max_{i=1}^d L_i$, this is a improvement in constant. When all L are equal, no improvement.

Steepest Coordinate Descent

- Algorithm Change (i) Choose $i=\underset{i \in[d]}{\operatorname{argmax}}\left|\nabla_i f\left(\mathbf{x}_t\right)\right|$, also called *Gauss-Southwell* rule.
- Since $\max_i\left|\nabla_i f(\mathbf{x})\right|^2 \geq \frac{1}{d} \sum_{i=1}^d\left|\nabla_i f(\mathbf{x})\right|^2$, for a coordinate-$L$-smooth function, we have
  - $f\left(\mathbf{x}_{t+1}\right) \leq f\left(\mathbf{x}_t\right)-\frac{1}{2 L} \max_i\left|\nabla_i f(\mathbf{x})\right|^2 \leq f\left(\mathbf{x}_t\right)-\frac{1}{2 L} \sum_{i=1}^d \frac{1}{d}\left|\nabla_i f\left(\mathbf{x}_t\right)\right|^2$
  - Then we arrive at Corollary 5.8, still $f\left(\mathbf{x}_T\right)-f\left(\mathbf{x}^{\star}\right) \leq\left(1-\frac{\mu}{d L}\right)^T\left(f\left(\mathbf{x}_0\right)-f\left(\mathbf{x}^{\star}\right)\right)$

Better Convergence of Steepest Descent with ℓ_1 norm strong convexity (Also known as Steeper)

- Lemma 5.9 (ℓ_1 norm μ-strong convexity -> ℓ_∞ norm μ-PL inequality) This is a natual resutl for *Lemma E*, as the dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$.
- Proof of dual norm $\|\cdot\|_{p *}=\|\cdot\|_q$ where $p^{-1}+q^{-1}=1$ is by Hölder inequality $\sum_{k=1}^n\left|x_k y_k\right| \leq\left(\sum_{k=1}^n\left|x_k\right|^p\right)^{\frac{1}{p}}\left(\sum_{k=1}^n\left|y_k\right|^q\right)^{\frac{1}{q}}$ (equality holds if x^p and y^q are proportional.)
  - Proof of special case of $(1, \infty)$ is simple.
- Theorem 5.10 Let $f: \mathbb{R}^d \rightarrow \mathbb{R} \in C^1$ has a global minimum $\mathbf{x}^{\star}$. Suppose that f is coordinate-wise L-smooth and satisfies the μ_1-PL inequality under ℓ_1 norm. Then *steepest coordinate descent* with step size $\gamma_i=L^{-1}$ arbitrary start satisfies $f\left(\mathbf{x}_T\right)-f\left(\mathbf{x}^{\star}\right) \leq\left(1-\frac{\mu_1}{L}\right)^T\left(f\left(\mathbf{x}_0\right)-f\left(\mathbf{x}^{\star}\right)\right)$.
  - Key Idea $f\left(\mathbf{x}_{t+1}\right) \leq f\left(\mathbf{x}_t\right)-\frac{1}{2 L} \max_i\left|\nabla_i f(\mathbf{x})\right|^2=f\left(\mathbf{x}_t\right)-\frac{1}{2 L}\left\|\nabla f\left(\mathbf{x}_t\right)\right\|_{\infty}^2$
- Comment Since $\|\mathbf{y}-\mathbf{x}\|_1 \geq\|\mathbf{y}-\mathbf{x}\| \geq\|\mathbf{y}-\mathbf{x}\|_1 / \sqrt{d}$
  - If f is ℓ_1 μ-strong convex -> f is ℓ_2 μ-strong convex.
  - If f is ℓ_2 μ-strong convex -> f is ℓ_1 μ/d-strong convex.
  - Seems no better than ℓ_2 case, but cases like $f(x)=x^{\top} \operatorname{diag}\left\{\lambda_i\right\} x / 2$ shows ℓ_1 gives much better results. (Appendic C of this)
  - Has sth to do with *convex conjugates*.

Greedy coordinate descent

- Algorithm Change (ii) $\mathbf{x}_{t+1}:=\underset{\lambda \in \mathbb{R}}{\operatorname{argmin}} f\left(\mathbf{x}_t+\lambda \mathbf{e}_i\right)$
  - Might be easier when f is analytic.
- Problem May stuck at non-minimum.
  - Example: $f(\mathbf{x}):=\|\mathbf{x}\|^2+\left|x_1-x_2\right|$, (x, x) is minimal for all single coordinate in range $|x| \leq 1/2$.
  - The following form is guaranteed to not be.
- Theorem 5.11 Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be of the form $f(\mathbf{x}):=g(\mathbf{x})+h(\mathbf{x})$ where $h(\mathbf{x})=\sum_i h_i\left(x_i\right)$, g, f_i all convex, $g \in C^1$, then whenever *greedy coordinate descent* makes no improvement, its in minimum.
  - Proof
    - No improvement means $\forall \lambda, i \in[d], f\left(\mathbf{x}+\lambda \mathbf{e}_i\right) \geq f(\mathbf{x})$,
      - then $f\left(\mathbf{x}+\left(y_i-x_i\right) \mathbf{e}_i\right) \geq f(\mathbf{x})$
      - let $p(z):=g\left(z, x_{-i}\right)-\partial_i g\left(x_i\right)\left(z-x_i\right)$ and $q(z):=h_i(z)+\partial_i g\left(x_i\right)\left(z-x_i\right)$, since g convex and differentiabl, $p(z)$ reach global minimum at $z=x_i$, $f_i:=f\left(z ; x_{-i}\right)=p(z)+q(z)$, so f_i also reach global minimum at $z=x_i$.
      - Since $p(z)$ differentiabl, $\partial p\left(x_i\right)=0$, if q does not reach global minimum at $z=x_i$, then $\exists y$ s.t. $q(y)<q\left(x_i\right)$, then by convexit $q\left(x_i+\lambda\left(y-x_i\right)\right) \leq q\left(x_i\right)+\lambda\left(q(y)-q\left(x_i\right)\right)$ is bounded by a negative slopt of $-\left(q\left(x_i\right)-q(y)\right)$ at $z=x_i$.
      - This leads to contradiction where $f=p+q$ reaches minimum at $z=x$.
      - Therefore, $q(z)$ also reaches minimal at $z=x_i$.
    - The above discussion means $\forall y_i, q\left(y_i\right) \geq q\left(x_i\right)$, equivalently $h_i\left(y_i\right)+\partial_i g(\mathbf{x})\left(y_i-x_i\right)-h\left(x_i\right) \geq 0$
    - Since g convex, $f(\mathbf{y}) \geq g(\mathbf{x})+\nabla g(\mathbf{x})^{\top}(\mathbf{y}-\mathbf{x})+h(\mathbf{y})-h(\mathbf{x})=f(\mathbf{x})+\sum_{i=1}^d \underbrace{\left(\nabla_i g(\mathbf{x})\left(y_i-x_i\right)+h_i\left(y_i\right)-h_i\left(x_i\right)\right)}_{\geq 0} \geq f(\mathbf{x})$
  - Comment LASSO $\min_{\mathbf{x} \in \mathbb{R}^n}\|A \mathbf{x}-\mathbf{b}\|^2+\lambda\|\mathbf{x}\|_1$ is of this form
    - Under mild regularity on g, convergence is affirmative.