Optimization for Data Science
ETH Zürich, FS 2023 261-5110-00L

Lecture 5: Coordinate Descent

Bernd Gärtner
Niao He

`https://www.ti.inf.ethz.ch/ew/courses/ODS23/index.html`

March 17, 2023

Motivation

Gradient descent:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$$

- computes and update d values in each iteration
- For large d, this can be problematic.

Coordinate descent: select some $i \in [d]$ and update only the i-th coordinate:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t)\mathbf{e}_i$$

- How do we choose the coordinate to update?
- Price to pay: more iterations?

Warmup: Alternative analysis of gradient descent...

...on smooth and strongly convex functions.

Before (Theorem 3.14): $\mathbf{x}_T$ converges to $\mathbf{x}^\star$ $\quad (\Rightarrow f(\mathbf{x}_T)$ converges to $f(\mathbf{x}^\star))$.

$$\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Now: $f(\mathbf{x}_T)$ converges to $f(\mathbf{x}^\star)$:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu}{L}\right)^T (f(\mathbf{x}_0)) - f(\mathbf{x}^\star))$$

For this, we can relax strong convexity. This allows to deal with

- several minimizers;
- even certain nonconvex functions!

The Polyak-Łojasiewicz inequality (1963)

Definition 5.1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with a global minimum $\mathbf{x}^\star$. We say that f satisfies the Polyak-Łojasiewicz inequality (PL inequality) if the following holds for some $\mu > 0$:

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f(\mathbf{x}^\star)), \quad \forall\, \mathbf{x} \in \mathbb{R}^d.$$

- Squared gradient norm at $\mathbf{x}$ is at least proportional to the error in objective function value at $\mathbf{x}$.
- Direct consequence: $\nabla f(\mathbf{x}) = \mathbf{0}$ (critical point) $\Rightarrow$ $\mathbf{x}$ is a global minimum.
- Strong convexity implies the PL inequality.

Strong convexity $\Rightarrow$ PL inequality

Lemma 5.2

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$ (in particular, a global minimum $\mathbf{x}^\star$ exists by Lemma 3.12). Then f satisfies the PL inequality for the same μ.

Proof.

$$
\begin{aligned}
f(\mathbf{x}^\star) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{x}^\star - \mathbf{x}) + \frac{\mu}{2}\left\|\mathbf{x}^\star - \mathbf{x}\right\|^2 \quad \text{(strong convexity)} \\
&\geq f(\mathbf{x}) + \min_{\mathbf{y}} \left(\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\left\|\mathbf{y} - \mathbf{x}\right\|^2 \right) \\
&= f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2.
\end{aligned}
$$

The PL inequality follows by simple rewriting. Last equation in the above proof:

- Solve for a critical point $\mathbf{y}^\star$ of the convex minimization problem.
- By Lemma 2.22, $\mathbf{y}^\star$ is a global minimum.

Strong convexity vs. PL inequality

The PL inequality is strictly weaker than strong convexity.

Example: $f(x_1, x_2) = x_1^2$

- Not strongly convex: every point $(0, x_2)$ is a global minimum.
- Satisfies the PL inequality in which it behaves like the strongly convex function $x \to x^2$, since gradient / function values do not depend on x_2.

There are even nonconvex functions satisfying the PL inequality (Exercise 35).

Gradient descent on smooth functions with PL inequality

Theorem 5.3

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$. Suppose that f is smooth with parameter L and satisfies the PL inequality with parameter $\mu > 0$. Choosing stepsize $\gamma = 1/L$, gradient descent with arbitrary $\mathbf{x}_0$ satisfies

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu}{L}\right)^T (f(\mathbf{x}_0)) - f(\mathbf{x}^\star)), \quad T > 0.$$

Proof.

For all t:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 && \text{(sufficient decrease, Lemma 3.7)} \\ &\leq f(\mathbf{x}_t) - \frac{\mu}{L}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) && \text{(PL inequality).} \end{aligned}$$

Subtract $f(\mathbf{x}^\star)$ on both sides:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)).$$

Coordinate-wise smoothness

A refined notion of smoothness that we can apply per coordinate.

Definition 5.4

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, and $\mathcal{L} = (L_1, L_2, \ldots, L_d) \in \mathbb{R}^d_+$. Function f is called coordinate-wise smooth (with parameter $\mathcal{L}$) if for every coordinate $i = 1, 2, \ldots, d$,

$$f(\mathbf{x} + \lambda \mathbf{e}_i) \leq f(\mathbf{x}) + \lambda \nabla_i f(\mathbf{x}) + \frac{L_i}{2}\lambda^2 \quad \forall \mathbf{x} \in \mathbb{R}^d, \lambda \in \mathbb{R}, .$$

If $L_i = L$ for all i, f is said to be coordinate-wise smooth with parameter L.

- If f is smooth with parameter L, then f is coordinate-wise smooth with parameter L. Proof: Apply standard smoothness inequality with $\mathbf{y} = \mathbf{x} + \lambda \mathbf{e}_i$.
- $f(x_1, x_2) = x_1^2 + 10x_2^2$ is smooth with $L = 20$ and coordinate-wise smooth with $\mathcal{L} = (2, 20)$.
- $f(x) = x_1^2 + x_2^2 + Mx_1x_2$ is smooth only with $L \geq (M+2)\sqrt{2}$ but coordinate-wise smooth with $L = 2$.

Coordinate descent algorithms

In Iteration t:

$$\text{choose some } i \in [d]$$
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t) \mathbf{e}_i .$$

- $\nabla_i f(\mathbf{x}_t)$ is the i-th entry of the gradient (i-th partial derivate).
- $\mathbf{e}_i$ is the i-th unit vector, so only the i-th coordinate of $\mathbf{x}_t$ is updated.
- γ_i is the stepsize for coordinate i.

Coordinate-wise sufficient decrease

Lemma5.5

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and coordinate-wise smooth with parameter $\mathcal{L} = (L_1, L_2, \ldots, L_d)$. With active coordinate i in iteration t and stepsize $\gamma_i = \frac{1}{L_i}$, coordinate descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2.$$

Proof.

Apply coordinate-wise smoothness with $\lambda = -\nabla_i f(\mathbf{x}_t)/L_i$ and $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda \mathbf{e}_i$.

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \lambda \nabla_i f(\mathbf{x}_t) + \frac{L_i}{2}\lambda^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L_i}|\nabla_i f(\mathbf{x}_t)|^2 + \frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2. \end{aligned}$$

$\square$

Randomized coordinate descent

sample $i \in [d]$ uniformly at random

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t)\mathbf{e}_i.$$

Nesterov [Nes12]:

- At least as fast as gradient descent on smooth functions, if it is d times cheaper to update one coordinate than the full iterate.

Karimi et al. [KNS16]:

- The same holds when we additionally assume the PL inequality.

Randomized coordinate descent: smooth functions, PL inequality

Theorem 5.6

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$. Suppose that f is coordinate-wise smooth with parameter L and satisfies the PL inequality with parameter $\mu > 0$. Choosing stepsize $\gamma_i = 1/L$ for all coordinates, randomized coordinate descent with arbitrary $\mathbf{x}_0$ satisfies

$$\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^\star)] \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star)), \quad T > 0.$$

Comparison with gradient descent:

- Number of iterations to reach error at most ε is by a factor of d higher.
- Follows from $(1 - \frac{\mu}{dL})$ vs. $(1 - \frac{\mu}{L})$.
- Zero-sum game: moved a factor of d from per-iteration complexity to iteration count.

Randomized coordinate descent: $\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^\star)] \le \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))$

Coordinate-wise sufficient decrease:

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}|\nabla_i f(\mathbf{x}_t)|^2.$$

Taking expectations with respect to the choice of the active coordinate i:

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_{t+1})|\mathbf{x}_t\right] &\le f(\mathbf{x}_t) - \frac{1}{2L}\sum_{i=1}^{d}\frac{1}{d}|\nabla_i f(\mathbf{x}_t)|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2dL}\|\nabla f(\mathbf{x}_t)\|^2 \quad \text{(Euclidean norm is very convenient)} \\
&\le f(\mathbf{x}_t) - \frac{\mu}{dL}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \quad \text{(PL inequality).}
\end{aligned}$$

Subtracting $f(\mathbf{x}^\star)$ from both sides:

$$\mathbb{E}[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star)|\mathbf{x}_t] \le \left(1 - \frac{\mu}{dL}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)).$$

Taking expectations with respect to $\mathbf{x}_t$:

$$\mathbb{E}[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star)] \le \left(1 - \frac{\mu}{dL}\right)\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)].$$

Importance sampling

Improves over uniform sampling when coordinate-wise smoothness parameters L_i differ.

sample $i \in [d]$ with probability $\dfrac{L_i}{\sum_{j=1}^d L_j}$

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L_i} \nabla_i f(\mathbf{x}_t) \mathbf{e}_i.$$

Theorem 5.7 (Exercise 36)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$, coordinate-wise smooth with parameter $\mathcal{L} = (L_1, L_2, \ldots, L_d)$, and satisfying the PL inequality with parameter $\mu > 0$. 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 and arbitrary $\mathbf{x}_0$ satisfies

$$\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^\star)] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star)), \quad T > 0.$$

Steepest coordinate descent

Deterministic algorithm, also known as the Gauss-Southwell rule:

$$\text{choose } i = \underset{i \in [d]}{\operatorname{argmax}} |\nabla_i f(\mathbf{x}_t)|$$

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_i \nabla_i f(\mathbf{x}_t) \mathbf{e}_i.$$

Corollary 5.8: Same number of iterations as randomized coordinate descent.

- Use $\max_i |\nabla_i f(\mathbf{x})|^2 \geq \frac{1}{d} \sum_{i=1}^{d} |\nabla_i f(\mathbf{x})|^2$.
- Do the analysis as for randomized coordinate descent, without expectations.

Iterations are more costly than in randomized coordinate descent, and we don't need less iterations. What's the point?

- We can still speed it up in some cases (next slide).
- Maximum absolute gradient may efficiently be maintainable throughout iterations.

Strong convexity with respect to ℓ_1-norm

Trick due to Nutini et al. [NSL+15]:

- Measure strong convexity w.r.t. ℓ_1-norm instead of ℓ_2-norm:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu_1}{2} \|\mathbf{y} - \mathbf{x}\|_1^2, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

- Then f is also strongly convex with $\mu = \mu_1$ in the usual sense.
  Proof: $\|\mathbf{y} - \mathbf{x}\|_1 \geq \|\mathbf{y} - \mathbf{x}\|$.
- If f is strongly convex with μ in the usual sense, then f is strongly convex with $\mu_1 = \mu/d$ w.r.t. ℓ_1-norm.
  Proof: $\|\mathbf{y} - \mathbf{x}\| \geq \|\mathbf{y} - \mathbf{x}\|_1 / \sqrt{d}$.
- $\mu \geq \mu_1 \geq \mu/d$.
- If $\mu_1 > \mu/d$, we can speed up steepest coordinate descent.

Strong convexity w.r.t. ℓ_1-norm $\Rightarrow$ PL inequality w.r.t. ℓ_∞-norm

Lemma 5.9 (Exercise 38)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu_1 > 0$ w.r.t. ℓ_1-norm. (In particular, f is μ_1-strongly convex w.r.t. Euclidean norm, so a global minimum $\mathbf{x}^\star$ exists by Lemma 3.12). Then f satisfies the PL inequality w.r.t. ℓ_∞-norm with the same μ_1:

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|_\infty^2 \geq \mu_1(f(\mathbf{x}) - f(\mathbf{x}^\star)), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

Same proof strategy as for the ℓ_2-norm / ℓ_2-norm case:

- Exercise 38: solve

$$\min_{\mathbf{y}} \left(\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu_1}{2}\|\mathbf{y} - \mathbf{x}\|_1^2 \right).$$

- This is still convex but non-differentiable, can't solve for a critical point.
- Elementary techniques apply (deeper reason why it works: convex conjugates).

Steeper (than steepest) coordinate descent

Theorem 5.10

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$. Suppose that f is coordinate-wise smooth with parameter L and satisfies the PL inequality w.r.t. ℓ_∞-norm with parameter $\mu_1 > 0$. Choosing stepsize $\gamma_i = 1/L$, steepest coordinate descent with arbitrary $\mathbf{x}_0$ satisfies

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu_1}{L}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star)), \quad T > 0.$$

Speedup?

- Normal steepest coordinate descent: $(1 - \frac{\mu}{dL})$.
- Worst case: $\mu_1 = \mu/d$, no speedup.
- Best case: $\mu_1 = \mu$, speedup by a factor of d.

Steeper coordinate descent: $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu_1}{L}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^\star))$

For all t:

Coordinate-wise sufficient decrease for $i = \mathrm{argmax}_{i \in [d]} |\nabla_i f(\mathbf{x}_t)|$:

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{1}{2L}|\nabla_i f(\mathbf{x}_t)|^2 = f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|_\infty^2 \\
&\leq f(\mathbf{x}_t) - \frac{\mu_1}{L}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)). \quad \text{(PL inequality w.r.t. } \ell_\infty\text{-norm)}
\end{aligned}
$$

Now it continues as for gradient descent (subtracting $f(\mathbf{x}^\star)$ from both sides):

$$
f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu_1}{L}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)),
$$

Greedy coordinate descent

Make the step that maximizes the progress in the chosen coordinate!

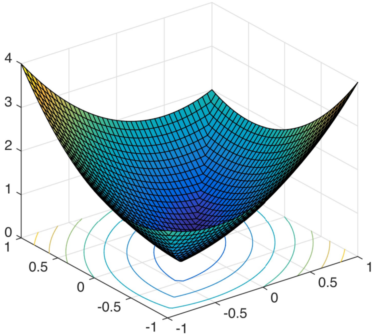
$$\text{choose } i \in [d]$$
$$\mathbf{x}_{t+1} := \underset{\lambda \in \mathbb{R}}{\operatorname{argmin}}\, f(\mathbf{x}_t + \lambda \mathbf{e}_i)$$

This requires to perform a line search.

- This can sometimes be done analytically, or approximately by some other means.
- Differentiable case: previous convergence bounds still hold as stepwise progress can only be better.
- Nondifferentiable case: algorithm may fail to converge!

Greedy coordinate descent failure

Example: $f(\mathbf{x}) := \|\mathbf{x}\|^2 + |x_1 - x_2|$.



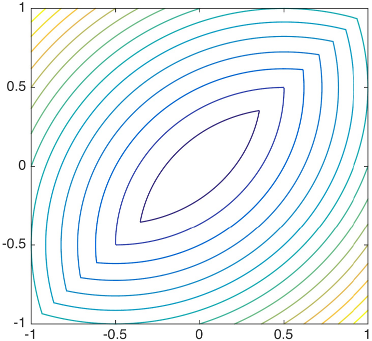


Figure by Alp Yurtsever & Volkan Cevher, EPFL

Global minimum is $(0, 0)$.

Greedy coordinate descent cannot escape any point $(x, x), |x| \leq 1/2$.

Saving greedy coordinate descent: the separable case

Theorem 5.11

Let $f : \mathbb{R}^d \to \mathbb{R}$ be of the form

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}) \quad \text{with } h(\mathbf{x}) = \sum_i h_i(x_i), \quad \mathbf{x} \in \mathbb{R}^d,$$

with g convex and differentiable, and the h_i convex. Let $\mathbf{x} \in \mathbb{R}^d$ be a point such that greedy coordinate descent cannot make progress in any coordinate. Then $\mathbf{x}$ is a global minimum of f.

A function h as in the theorem is called separable.

Popular examples: regularizers $h(\mathbf{x}) = \|\mathbf{x}\|_1$ and $h(\mathbf{x}) = \|\mathbf{x}\|^2$.

Convergence of greedy coordinate descent does not automatically follow but can be proved (under mild conditions) [Tse01].

Example: LASSO, Lagrange dual version

LASSO with tuning parameter R:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{w}) = \sum_{i=1}^n \|\mathbf{w}^\top \mathbf{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathbf{w}\|_1 \leq R, \end{array}$$

Lagrange dual function $g(\lambda), \lambda \geq 0$:

$$\text{minimize} \quad F_\lambda(\mathbf{w}) = f(\mathbf{w}) + \lambda(\|\mathbf{w}\|_1 - R)$$

If $n \geq d$, we can assume that f (and hence F_λ) are strictly convex, so the LASSO solution $\mathbf{w}^\star$ and the dual solutions $\mathbf{w}(\lambda)$ are unique.

- LASSO is a convex program with a Slater point, so by Theorem 2.48, there is $\lambda^\star \geq 0$ such that—using complementary slackness in the first equation:

$$F_{\lambda^\star}(\mathbf{w}^\star) = f(\mathbf{w}^\star) = g(\lambda^\star) = \min_{\mathbf{w}} F_{\lambda^\star}(\mathbf{w}) = F_{\lambda^\star}(\mathbf{w}(\lambda^\star)) \quad \Rightarrow \mathbf{w}^\star = \mathbf{w}(\lambda^\star).$$

- Hence, $\mathbf{w}^\star$ is also a minimizer of $f(\mathbf{w}) + \lambda^*\|\mathbf{w}\|_1$, but λ^* is unknown.
- LASSO, dual version: minimize $f(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$ with tuning parameter λ.
- $f(\mathbf{w})$ is convex and differentiable, $\lambda\|\mathbf{w}\|_1$ nondifferentiable but separable.

Summary

Coordinate descent methods are used widely in machine learning.

- State of the art for generalized linear models, including linear classifiers and regression models, with separable convex regularizers (e.g. ℓ_1-norm or squared ℓ_2-norm).

Results on coordinate-wise smooth and strongly convex functions (we only need the PL inequality, a consequence of strong convexity):

Algorithm	PL norm	Smoothness	Bound	Result
Randomized	ℓ_2	L	$1 - \frac{\mu}{dL}$	Theorem 5.6
Importance sampling	ℓ_2	$(L_1, L_2, \ldots, L_d)$	$1 - \frac{\mu}{d\bar{L}}$	Theorem 5.7
Steepest	ℓ_2	L	$1 - \frac{\mu}{dL}$	Corollary 5.8
Steeper (than Steepest)	ℓ_1	L	$1 - \frac{\mu_1}{L}$	Theorem 5.10

In the worst case, nothing is gained over gradient descent, and Steepest may even lose.

In the best case, Importance sampling and Steeper (than Steepest) may be up to d times faster than gradient descent.

Bibliography

Hamed Karimi, Julie Nutini, and Mark Schmidt.
Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition.
In *ECML PKDD 2016: Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Yurii Nesterov.
Efficiency of coordinate descent methods on huge-scale optimization problems.
SIAM Journal on Optimization, 22(2):341–362, 2012.

Julie Nutini, Mark W Schmidt, Issam H Laradji, Michael P Friedlander, and Hoyt A Koepke.
Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection.
In *ICML - Proceedings of the 32nd International Conference on Machine Learning*, pages 1632–1641, 2015.

P. Tseng.
Convergence of a block coordinate descent method for nondifferentiable minimization.
Journal of Optimization Theory and Applications, 109(3):475–494, 2001.