Chapter 9

Coordinate Descent

Contents

9.1 Coordinate Descent

Coordinate descent methods generate a sequence $\{\mathbf{x}_t\}_{t\geq 0}$ of iterates as follows:

$$\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma \mathbf{e}_{i_t} \,, \tag{9.1}$$

where $\mathbf{e}_i$ denotes the i-th unit basis vector in $\mathbb{R}^d$, and γ is a suitable stepsize for the selected coordinate of our objective function. Here we will focus on the gradient-based choice of the stepsize as

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \tfrac{1}{L} \nabla_{i_t} f(\mathbf{x}_t)\, \mathbf{e}_{i_t} \,, \tag{9.2}$$

As an alternative, for some problems we can find an even better stepsize by solving the single-variable minimization $\operatorname{argmin}_{\gamma\in\mathbb{R}} f(\mathbf{x}_t + \gamma \mathbf{e}_{i_t})$ in closed form.

9.2 Randomized Coordinate Descent

In *random coordinate descent*, the active coordinate i_t in each step is chosen uniformly at random from the set $[d]$.

[Nes12] shows that randomized coordinate descent achieves a faster convergence rate than gradient descent, if our problem of interest has d variables and it is assumed to be d times cheaper to update one coordinate than it is to compute the full gradient.

Convergence Analysis. To analyze coordinate descent methods, we assume *coordinate-wise smoothness* of f, which is defined as

$$f(\mathbf{x} + \gamma \mathbf{e}_i) \leq f(\mathbf{x}) + \gamma \nabla_i f(\mathbf{x}) + \frac{L}{2}\gamma^2 \quad \forall \mathbf{x} \in \mathbb{R}^d,\ \forall \gamma \in \mathbb{R}, \tag{9.3}$$

for any coordinate i. As with our familiar definition of smoothness, the property here is equivalent to the gradient being coordinate-wise Lipschitz continuous, that is $|\nabla_i f(\mathbf{x} + \gamma \mathbf{e}_i) - \nabla_i f(\mathbf{x})| \leq L|\gamma|,\ \ \forall \mathbf{x} \in \mathbb{R}^d, \gamma \in \mathbb{R}, i \in [d]$. We have seen the equivalence in Lemma 2.7 previously.

If we additionally assume strong convexity, we can obtain a fast linear convergence rate as follows.

Theorem 9.1. *Consider minimization of a function f which is coordinate-wise smooth with constant L as in (9.3), and is strongly convex with parameter $\mu > 0$. Then, coordinate descent with a stepsize of $1/L$,*

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \tfrac{1}{L}\nabla_{i_t} f(\mathbf{x}_t)\, \mathbf{e}_{i_t} \,.$$

when choosing the active coordinate i_t uniformly at random, has an expected linear convergence rate of

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \le \left(1 - \frac{\mu}{dL}\right)^t [f(\mathbf{x}_0) - f^\star].$$

Proof. We follow [KNS16]. By pluggin the update rule (9.1) into the smoothness condition (9.3) we have the step improvement

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

By taking the expectation of both sides with respect to i_t we have

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] &\le f(\mathbf{x}_t) - \frac{1}{2L}\mathbb{E}\left[|\nabla_{i_t} f(\mathbf{x}_t)|^2\right] \\
&= f(\mathbf{x}_t) - \frac{1}{2L}\frac{1}{d}\sum_i |\nabla_i f(\mathbf{x}_t)|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2dL}\|\nabla f(\mathbf{x}_t)\|^2.
\end{aligned}$$

We now use the the fact that strongly convex functions satisfy $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \ge \mu(f(\mathbf{x}) - f^\star) \ \ \forall\, \mathbf{x}$. This is proven in Lemma 9.2 below and is a property of separate interest. Subtracting $f^\star$ from both sides, we therefore obtain

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

Applying this recursively and using iterated expectations yields the result. □

For the algorithm variant using exact coordinate optimization instead of using the fixed stepsize $1/L$, the same result still holds (since progress per step is at least as good).

9.2.1 The Polyak-Łojasiewicz Condition

A function f satisfies the *Polyak-Łojasiewicz Inequality* (PL) if the following holds for some $\mu > 0$,

$$\tfrac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^\star), \quad \forall\, \mathbf{x}. \tag{9.4}$$

The condition was proposed by Polyak in 1963, and also by Łojasiewicz in the same year. It implies the quadratic growth condition.

Lemma 9.2 (Strong Convexity $\Rightarrow$ PL)**.** *Let f be strongly convex with parameter $\mu > 0$. Then f satisfies PL for the same μ.*

Proof. For all $\mathbf{x}$ and $\mathbf{y}$ we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 .$$

minimizing each side of the inequality with respect to $\mathbf{y}$ we obtain

$$f(\mathbf{x}^\star) \geq f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2,$$

which implies the PL inequality holds with the same value μ. □

The PL condition is a weaker condition than strong convexity. For example, it can be shown that it is satisfied for all compositions $f(\mathbf{x}) := g(A\mathbf{x})$ for strongly convex g and arbitrary matrix A, including least squares regression and many other applications in machine learning.

As we have seen in the proof of the above theorem, the linear convergence rate holds not only for strongly convex objectives but indeed for the wider class of any f satisfying the PL condition:

Corollary 9.3. *For minimization of a function f which is coordinate-wise smooth with constant L as in (9.3), satisfies the PL inequality (9.4), and has a non-empty solution set $\mathcal{X}^\star$, random coordinate descent with a stepsize of $1/L$ has the expected linear convergence rate of*

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \leq \left(1 - \frac{\mu}{dL}\right)^t [f(\mathbf{x}_0) - f^\star].$$

Using the same proof technique, gradient descent can be shown to exhibit a linear convergence rate for PL functions as well, see Exercise 38.

9.2.2 Importance Sampling

Uniformly random selection of the active coordinate might not always be the best choice. Let us consider an individual smoothness constant L_i for each coordinate i, that is

$$f(\mathbf{x} + \gamma \mathbf{e}_i) \leq f(\mathbf{x}) + \gamma \nabla_i f(\mathbf{x}) + \tfrac{L_i}{2}\gamma^2 \tag{9.5}$$

for all $\mathbf{x} \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$. In this case, instead of uniform random sampling of the active coordinate, it makes sense to sample proportional to the L_i values as suggested by [Nes12]. Formally, the selection rules picks i with probability $P[i_t = i] = \frac{L_i}{\sum_i L_i}$.

For coordinate descent using this modified sampling probabilities, and using a stepsize of $1/L_{i_t}$, the same convergence argument as above can be shown (Exercise 39) to give the faster rate of

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^t [f(\mathbf{x}_0) - f^\star],$$

where $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$ now is the average of all coordinate-wise smoothness constants. Note that this value can be much smaller than the global L we have used above, since that one was required to hold for all i so has to be chosen as $L = \max_i L_i$ instead.

Similar importance sampling strategies work for the different setting of stochastic gradient descent (SGD) on sum-structured problems. Practical performance of importance sampling over uniform sampling can be very significant for coordinate descent (or SGD) in particular for sparse or inhomogeneous data.

9.3 Steepest Coordinate Descent

In contrast to random coordinate descent, *steepest coordinate descent* (or greedy coordinate descent) chooses the active coordinate according to

$$i_t := \underset{i \in [d]}{\operatorname{argmax}} \left| \nabla_i f(\mathbf{x}_t) \right|. \tag{9.6}$$

which is also called the Gauss-Southwell (GS) rule.

Convergence Analysis. It is easy to show that the same convergence rate which we have obtained for random coordinate descent in Theorem 9.1 also holds for steepest coordinate descent. To see this, the only ingredient we need is the fact that

$$\max_i |\nabla_i f(\mathbf{x})|^2 \geq \frac{1}{d} \sum_i |\nabla_i f(\mathbf{x})|^2 ,$$

and since we now have a deterministic algorithm, there is no need to take expectations in the proof.

Corollary 9.4. *For minimization of a function f which is coordinate-wise smooth with constant L as in (9.3), and is strongly convex with parameter $\mu > 0$, steepest coordinate descent with a stepsize of $1/L$ has the linear convergence rate of*

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \leq \left(1 - \frac{\mu}{dL}\right)^t [f(\mathbf{x}_0) - f^\star].$$

It was shown by [NSL+15] that a stronger convergence result can be obtained for this algorithm when the strong convexity of f is measured with respect to the ℓ_1-norm instead of the standard Euclidean norm, i.e.

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu_1}{2} \|\mathbf{y} - \mathbf{x}\|_1^2 .$$

Theorem 9.5. *For minimization of a function f which is coordinate-wise smooth with constant L as in (9.3), and is strongly convex w.r.t. the ℓ_1-norm with parameter $\mu_1 > 0$, steepest coordinate descent with a stepsize of $1/L$ has the linear convergence rate of*

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \leq \left(1 - \frac{\mu_1}{L}\right)^t [f(\mathbf{x}_0) - f^\star].$$

The proof again directly follows the one of Theorem 9.1, but uses the following lemma measuring the PL inequality in the ℓ_∞-norm:

Lemma 9.6. *Let f be strongly convex w.r.t. the ℓ_1-norm with parameter $\mu_1 > 0$. Then f satisfies*

$$\tfrac{1}{2} \|\nabla f(\mathbf{x})\|_\infty^2 \geq \mu_1 (f(\mathbf{x}) - f^\star).$$

The proof of the lemma is not given here, but follows the same strategy as in the earlier analogue Lemma 9.2. It then uses a property of convex conjugate functions (coming from the fact that the norms $\|.\|_1$ and $\|.\|_\infty$ are dual to each other).

In summary, we have that steepest coordinate descent can be up to d times faster than random coordinate descent in terms of number of iterations. However, of course the selection rule is now more costly. Naively, finding the steepest coordinate would require computing the full gradient, and might also cost d times more than using a random coordinate.

Steepest coordinate descent is nevertheless an attractive choice for problem classes where we can obtain (or maintain) the steepest coordinate efficiently. This includes several practical case, for example when the gradients are sparse, e.g. because the original data is sparse. Another important use-case is for problems where we would want to find a solution in as few steps as possible, i.e. a sparse solution. For example, the Lasso problem is interesting in terms of both mentioned aspects. Last but not least, we note that the steepest selection rule (9.6) looks very similar to the Frank-Wolfe algorithm, if one is optimizing over an ℓ_1-ball. This is not a coincidence, but indeed the two algorithms and their convergence are closely related in that case.

9.4 Non-smooth objectives

So far, we have only considered unconstrained and smooth optimization problems in this chapter.

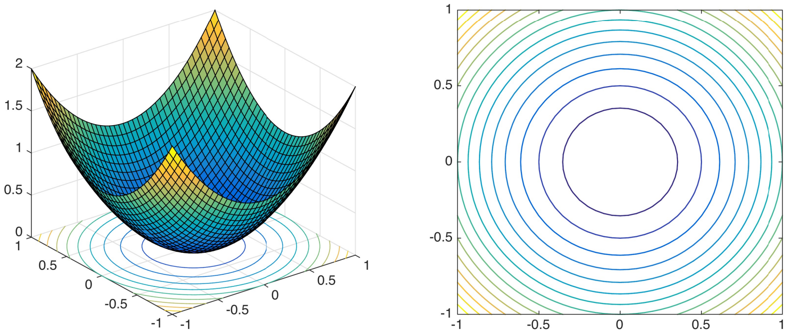

Figure 9.1: A smooth function: $f(\mathbf{x}) := \|\mathbf{x}\|^2$.

We have just proven that coordinate decent converges for differentiable, smooth f. What if f is not differentiable at all points? Earlier, when we analyzed gradient methods, we saw that in that case the extension to subgradients was straightforward and maintained the convergence results up to a small slowdown. Unfortunately for coordinate descent, the situation is not that easy.

Even when using exact minimization on each coordinate step, the algorithm can get permanently stuck in non-optimal points, as for example shown in the objective function of Figure 9.2: Not all hope is lost however. Consider the class of composite problems (recall proximal gradient descent as we discussed in Section 3.6),

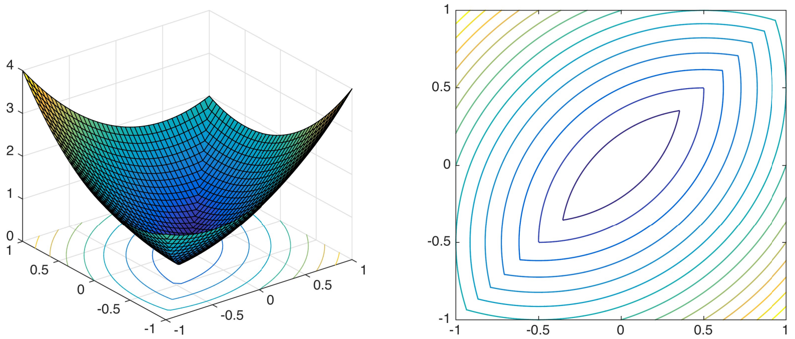

Figure 9.2: A non-smooth function: $f(\mathbf{x}) := \|\mathbf{x}\|^2 + |x_1 - x_2|$.

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}) \quad \text{with } h(\mathbf{x}) = \sum_i h_i(x_i)\,, \tag{9.7}$$

for g convex and smooth, and $h(\mathbf{x}) = \sum_i h_i(x_i)$ separable with h_i convex but possibly non-smooth. For this class of problems, coordinate descent with exact minimization converges to a global optimum, as illustrated in Figure 9.3.

One very important class of applications here are smooth functions f combined with ℓ_1-regularization, such as the Lasso.

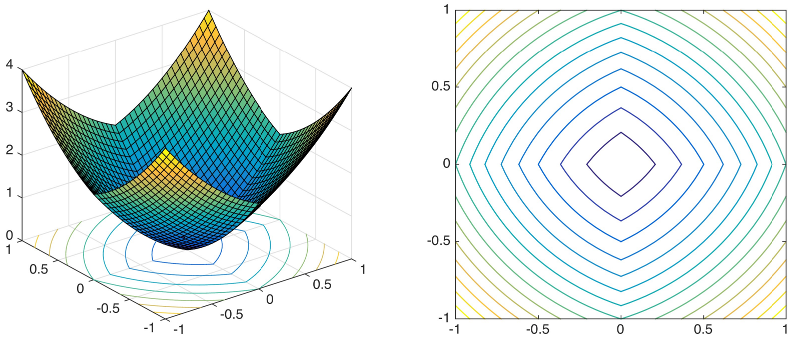

Figure 9.3: A function with separable non-smooth part: $f(\mathbf{x}) := \|\mathbf{x}\|^2 + \|\mathbf{x}\|_1$.

9.5 Applications

Coordinate descent methods are used widely in classic machine learning applications. Variants of coordinate methods form the state of the art for the class of generalized linear models, including linear classifiers and regression models, as long as separable convex regularizers are used (e.g. ℓ_1 or ℓ_2 norm regularization).

For least-squares linear regression $f(\mathbf{x}) := \|A\mathbf{x} - \mathbf{b}\|^2$, exact coordinate minimization can easily be performed readily in closed form.

Lasso. The optimization problem for sparse least squares linear regression (also known as the Lasso) is given by

$$\min_{\mathbf{x} \in \mathbb{R}^n} \; \|A\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{x}\|_1 \tag{9.8}$$

for some regularization parameter $\lambda > 0$. It is an instance of our class of composite optimization problems (9.7).

Support Vector Machines. The original optimization problem for the Support Vector Machine (SVM) is given by

$$\min_{\mathbf{w} \in \mathbb{R}^d} \; \sum_{i=1}^{n} \ell(y_i A_i^\top \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{9.9}$$

where $\ell : \mathbb{R} \to \mathbb{R}$, $\ell(z) := \max\{0, 1 - z\}$ is the *hinge loss* function. Here for any i, $1 \le i \le n$, the vector $A_i \in \mathbb{R}^d$ is the i-th data example, and $y_i \in \{\pm 1\}$ is the corresponding label.

The dual optimization problem for the SVM is given by

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \ \boldsymbol{\alpha}^\top \mathbf{1} - \tfrac{1}{2\lambda} \boldsymbol{\alpha}^\top Y A^\top A Y \boldsymbol{\alpha} \quad \text{such that} \quad 0 \le \alpha_i \le 1 \ \forall i \tag{9.10}$$

where $Y := diag(\mathbf{y})$, and $A \in \mathbb{R}^{d \times n}$ again collects all n data examples as its columns. The dual problem is an instance of our class of composite optimization problems (9.7), since the non-differentiable box-constraint $0 \le \alpha_i \le 1 \ \forall i$ can be written as a separable g as required.

9.6 Exercises

Exercise 38 (Alternative analysis for gradient descent)**.** *Let f be smooth with constant L in the classical sense, and satisfy the PL inequality (9.4). Let the problem $\min_{\mathbf{x}} f(\mathbf{x})$ have a non-empty solution set $\mathcal{X}^\star$. Prove that* gradient descent *with a stepsize of $1/L$ has a global linear convergence rate*

$$f(\mathbf{x}_t) - f^\star \le \left(1 - \frac{\mu}{L}\right)^t (f(\mathbf{x}_0) - f^\star).$$

Exercise 39 (Importance Sampling)**.** *Consider random coordinate descent with selecting the i-th coordinate with probability proportional to the L_i value, where L_i is the individual smoothness constant for each coordinate i as in (9.5).*

When using a stepsize of $1/L_{i_t}$, prove that we obtain the faster rate of

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \le \left(1 - \frac{\mu}{d\bar{L}}\right)^t [f(\mathbf{x}_0) - f^\star],$$

where $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$ now is the average of all coordinate-wise smoothness constants. Note that this value can be much smaller than the global L we have used above, since that one was required to hold for all i so has to be chosen as $L = \max_i L_i$ instead.

Can you come up with an example from machine learning where $\bar{L} \ll L$?

Exercise 40. *Derive the solution to exact coordinate minimization for the Lasso problem (9.8), for the i-th coordinate. Write A_{-i} for the $(d-1) \times n$ matrix obtained by removing the i-th column from A.*

Bibliography

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. https://web.stanford.edu/~boyd/cvxbook/.

[Dav59] William C. Davidon. Variable metric method for minimization. Technical Report ANL-5990, AEC Research and Development, 1959.

[Dav91] William C. Davidon. Variable metric method for minimization. *SIAM J. Optimization*, 1(1):1–17, 1991.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 07 2008.

[Gol70] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[Gre70] J. Greenstadt. Variations on variable-metric methods. *Mathematics of Computation*, 24(109):1–22, 1970.

[KNS16] 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 International Publishing, Cham, September 2016.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[Noc80] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006.

[NSL$^+$15] 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*, pages 1632–1641, 2015.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[Vis14] Nisheeth Vishnoi. Lecture notes on fundamentals of convex optimization, 2014. `https://tcs.epfl.ch/files/content/sites/tcs/files/Lec3-Fall14-Web.pdf`.

[Zim16] Judith Zimmermann. *Information Processing for Effective and Stable Admission*. PhD thesis, ETH Zurich, 2016. .