Optimization in Machine Learning and Statistics

Lecture notes, 3A MIR

Yohann De Castro
École Centrale de Lyon

January 2027

Abstract

These notes cover convex optimization for Machine Learning and High-Dimensional Statistics. They present the geometry of convexity, duality and the KKT conditions, first-order methods (gradient descent, stochastic gradient, subgradient, coordinate descent, Frank-Wolfe), proximal algorithms and operator splitting (ISTA, FISTA, Douglas-Rachford, ADMM, Chambolle-Pock), with the monotone operator tools they use, and second-order methods (Newton, quasi-Newton, interior point methods). Each result is either proved in full or marked as admitted, with a reference.

Contents

Introduction 4

1 The Geometry of Convexity 7
1.1 Convex Sets . . . 7
1.2 Bridging Sets and Functions: The Epigraph . . . 8
1.3 Convex Functions: Analysis and Characterization . . . 9
1.4 Subgradients . . . 10
1.5 Curvature Refined: Smoothness and Strong Convexity . . . 11
1.6 Relaxing Convexity: The Polyak-Lojasiewicz (PL) Condition . . . 12
1.7 Duality and Optimality . . . 14

2 First-Order Methods 17
2.1 Gradient Descent (GD) . . . 17
2.2 Stochastic Gradient Descent (SGD) . . . 18
2.3 Variance Reduction (SVRG) . . . 21
2.4 The Subgradient Method . . . 21
2.5 Adaptive Methods: Adam and AdamW . . . 22
2.6 Coordinate Descent (CD) . . . 23
2.7 Screening Rules (SAFE) . . . 23
2.8 Frank-Wolfe (Conditional Gradient) . . . 24

3 Proximal Algorithms and Operator Splitting Methods 26
3.1 The Proximal Operator: A Majorization Perspective . . . 26
3.2 ISTA: Iterative Shrinkage-Thresholding a.k.a. Forward-Backward (FB) . . . 27
3.3 FISTA (Fast ISTA) and Nesterov Acceleration . . . 29
3.4 The Proximal Point Algorithm . . . 30
3.5 Monotone Operators and the Optimization Connection . . . 30
3.6 Douglas-Rachford Splitting . . . 32
3.7 ADMM: Alternating Direction Method of Multipliers . . . 33
3.8 Linear Convergence of Splitting Methods . . . 33
3.9 Equivalence of ADMM and Dual Douglas-Rachford . . . 36
3.10 Primal-Dual Splitting (Chambolle-Pock) . . . 37
3.11 Accelerated Chambolle-Pock for Strongly Convex Objectives . . . 39

4 Second-Order Methods 41
4.1 Newton's Method . . . 41
4.2 Quasi-Newton Methods (BFGS) . . . 42
4.3 Interior Point Methods (IPM) . . . 43

5 Numerical Experiments 44
5.1 Benchmark 1: Ridge Regression . . . 44
5.2 Benchmark 2: LASSO . . . 45

6 Summary of Main Results 47

List of Notation 50

Glossary 52

List of Figures

1 Linear trend filtering (ℓ_1 penalty on second differences), $n = 100$ equispaced points, noise standard deviation 0.15, $\lambda = 4$. The true signal is piecewise linear with knots at 0.3, 0.55 and 0.8. The solution has four knots, at 0.303, 0.535, 0.556 and 0.808: the true knot at 0.55 is rendered by two neighboring knots. Session 1 notebook, experiment 2. 5
2 Matrix completion: we observe a subset of the entries (for instance, movie ratings) and infer the others using a low-rank assumption. 5
3 The Separating Hyperplane Theorem. The vector a is normal to the hyperplane separating C and D. 7
4 The vector $(g, -1)$ is normal to the supporting hyperplane of the epigraph at $(x, f(x))$; here f is differentiable and $g = f'(x)$. 9
5 Conditioning: an L-smooth and m-strongly convex function (black) lies between the wide parabola defined by m (blue) and the narrow parabola defined by L (red), both tangent to f at x; this holds at every base point x. 11
6 Schematic picture of momentum: on an ill-conditioned problem, GD (red) may oscillate across the narrow valley, while momentum methods (blue) average successive directions. This is an illustration, not a computed trajectory. 18
7 Safe screening (schematic): if a ball $\mathcal{B}$ known to contain the dual solution $\theta^\star$ does not meet the hyperplanes $H_j = \{\theta : |X_j^T \theta| = \lambda\}$, then $\beta_j^\star = 0$. 23
8 Geometry of a saddle point problem, drawn for the model $\mathcal{L}(x, y) = x^2 - y^2$ (schematic). *(a)* Level curves of $\mathcal{L}$ (hyperbolas; blue: positive levels, red: negative levels) and a schematic primal-dual trajectory (purple), not computed. *(b)* One-dimensional cross-sections: fixing $y = y^\star$ gives a convex function of x (blue), fixing $x = x^\star$ gives a concave function of y (red). The saddle point minimizes the first and maximizes the second. 38
9 Interior point methods (schematic): instead of moving along the boundary like the simplex method, IPMs follow the central path through the interior. 43
10 Ridge regression on synthetic data ($n = 20\,640$, $d = 8$, $\kappa = 110.4$): gap $f(w) - f^\star$ against epochs. Gaps are clipped at 10^{-16} for the log scale. 44
11 LASSO on the same synthetic data, $\lambda = \lambda_{\max}/10$: gap $F(w) - F^\star$ against iterations (for CD, one sweep over the coordinates). Gaps are clipped at 10^{-16} for the log scale. 45

Introduction

Machine learning has two sides: the statistical task of modeling dependencies (the *inverse problem*) and the computational task of finding the parameters that best fit these models. Typically, we assume a generative model $y \approx \Phi(x)$, where x is the signal and y is the observation. We recover x by solving:

$$(P): \quad \min_{x \in \mathcal{D}} \quad \underbrace{f(x)}_{\text{Data Fidelity}} \quad + \quad \underbrace{\lambda \mathcal{R}(x)}_{\text{Regularization}} \tag{0.1}$$

where f measures the divergence between prediction and observation, and $\mathcal{R}$ enforces structural priors (e.g., sparsity, low-rank, smoothness). While historical taxonomies distinguished between Linear and Non-Linear programming, the main divide today is between *convex* and *non-convex* optimization. This distinction matters: for a convex problem, every local minimum is global (Section 1.3).

Motivating Example 1: Linear Trend Filtering

Consider a non-parametric regression scenario where we observe a sequence of noisy data points $y_1, \dots, y_n \in \mathbb{R}$ corresponding to equispaced time intervals $t_1, \dots, t_n$. The statistical objective is to recover an underlying signal $\theta \in \mathbb{R}^n$ that filters out the noise while preserving the significant structural shifts in the trend.

A classical approach might employ a moving average or Ridge regression, both of which smooth the signal indiscriminately, blurring sharp transitions. *Linear Trend Filtering* takes a different approach. It seeks a piecewise linear estimator with adaptively chosen knots (change points). This is formulated as the following convex optimization problem:

$$\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-2} |\theta_i - 2\theta_{i+1} + \theta_{i+2}| \right\}. \tag{0.2}$$

The first term is the least-squares data fidelity term. The second term is an ℓ_1 penalty on the discrete second difference of θ. The term $(\theta_i - 2\theta_{i+1} + \theta_{i+2})$ is, up to the factor $(t_{i+1} - t_i)^2$, a discrete approximation of the second derivative θ'' at time t_{i+1}. Penalizing the ℓ_1 norm of these differences favors vectors with many zero second differences.

A function with a zero second derivative is linear; therefore, a function with a sparse second derivative is piecewise linear. This problem is a specific instance of the *Generalized Lasso*:

$$\min_{\theta} \left\{ \frac{1}{2} \|y - \theta\|_2^2 + \lambda \|D^{(k+1)} \theta\|_1 \right\}.$$

where $D^{(k+1)}$ is the discrete difference operator of order $k+1$. This estimator is related to regression splines . For $k = 0$, this is Total Variation (TV) denoising (fused Lasso), whose solutions are piecewise constant. For $k = 1$, as above, we obtain piecewise linear trends. In the PGD course, TV is the total variation norm of a measure; on a grid, it is the ℓ_1 norm of the weights.

Motivating Example 2: Matrix Completion

The second motivating example arises from the domain of recommender systems and collaborative filtering. We are presented with a partially observed matrix $Y \in \mathbb{R}^{m \times n}$ where entries are known only on a subset of indices $\Omega \subset \{1, \dots, m\} \times \{1, \dots, n\}$. The goal is to impute the missing entries. Without further assumptions, this problem is ill-posed; any completion is valid. However, we often assume the matrix is *low-rank*: user preferences are determined by a small number of latent factors.

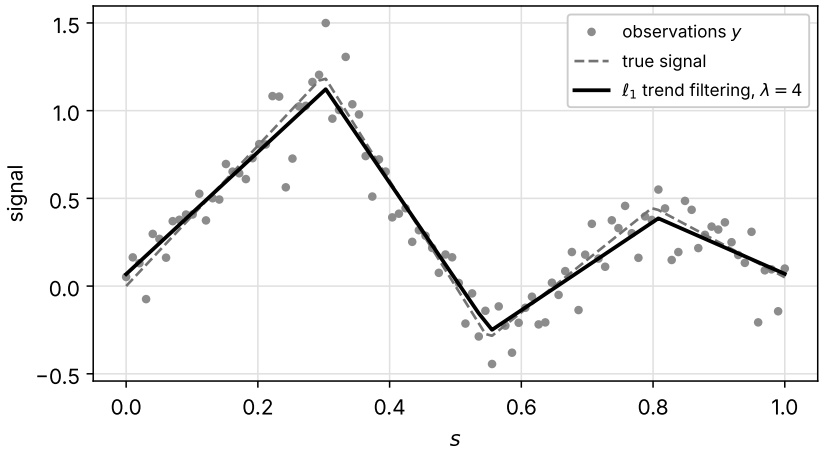
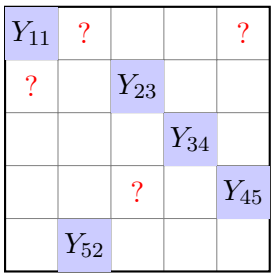


Figure 1: Linear trend filtering (ℓ_1 penalty on second differences), $n = 100$ equispaced points, noise standard deviation 0.15, $\lambda = 4$. The true signal is piecewise linear with knots at 0.3, 0.55 and 0.8. The solution has four knots, at 0.303, 0.535, 0.556 and 0.808: the true knot at 0.55 is rendered by two neighboring knots. Session 1 notebook, experiment 2.



Matrix Completion
Given observations on Ω (blue),
recover the full low-rank matrix X.
Problem: $\min \|X\|_*$ s.t. $P_\Omega(X) = P_\Omega(Y)$

Figure 2: Matrix completion: we observe a subset of the entries (for instance, movie ratings) and infer the others using a low-rank assumption.

A first formulation seeks the matrix of minimum rank consistent with the observed data:

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{s.t.} \quad X_{ij} = Y_{ij}\,, \ \forall (i,j) \in \Omega$$

Minimizing the rank, a non-convex function with integer values, under such constraints is NP-hard in general. A standard relaxation replaces the rank with the *nuclear norm* $\|X\|_*$, which is the convex envelope of the rank on the set $\{X : \|X\|_{op} \leq 1\}$:

$$\min_X \|X\|_* \quad \text{s.t.} \quad X_{ij} = Y_{ij}\,, \ \forall (i,j) \in \Omega \tag{0.3}$$

The nuclear norm is the sum of the singular values of X, i.e., $\|X\|_* = \sum_i \sigma_i(X)$. Under an incoherence condition (the singular vectors are not too aligned with the canonical basis) and for observed entries drawn at random, nuclear norm minimization recovers the low-rank matrix exactly with high probability .

Motivating Example 3: Deep Learning

The previous examples admit convex formulations (or relaxations). Training a neural network is a non-convex problem. A neural network is a composition of parameterized functions:

$$\Phi_\theta(x) = W_L \sigma(\dots \sigma(W_1 x) \dots) \tag{0.4}$$

where σ is a non-linear activation (e.g., ReLU). The optimization problem becomes:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\Phi_\theta(x_i), y_i) \tag{0.5}$$

Unlike the Lasso or matrix completion, this objective is non-convex in general, and it may have saddle points and non-global local minima. Why first-order methods such as SGD (Section 2) work well on these problems is not fully understood.

1 The Geometry of Convexity

This section introduces convex sets and functions, the curvature constants that govern the speed of algorithms, and the duality theory used throughout the notes.

1.1 Convex Sets

Definition 1.1 (Convex Set)**.** A set $C \subseteq \mathbb{R}^n$ is convex if for any $x, y \in C$ and any $\theta \in [0,1]$, the convex combination lies in C:

$$\theta x + (1-\theta)y \in C$$

In words, a convex set contains the segment between any two of its points.

Remark 1.2. The intersection of any collection (finite or infinite) of convex sets is convex. This is often the simplest way to prove convexity. For instance, a polyhedron $P = \{x \mid Ax \leq b\}$ is the intersection of finitely many halfspaces, hence it is convex. The cone of positive semidefinite (PSD) matrices is the intersection of the halfspaces $\{X : z^T X z \geq 0\}$ over all vectors z, hence it is convex.

Separation Theorems

The separating hyperplane theorem is the geometric tool behind duality theory and the KKT conditions: two disjoint convex sets lie on the two sides of a hyperplane.

Theorem 1.3 (Separating Hyperplane Theorem)**.** *Let $C, D \subseteq \mathbb{R}^n$ be two disjoint, nonempty convex sets. Then there exists a nonzero vector $a \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}$ such that:*

$$a^T x \leq b \quad \forall x \in C \quad \text{and} \quad a^T y \geq b \quad \forall y \in D$$

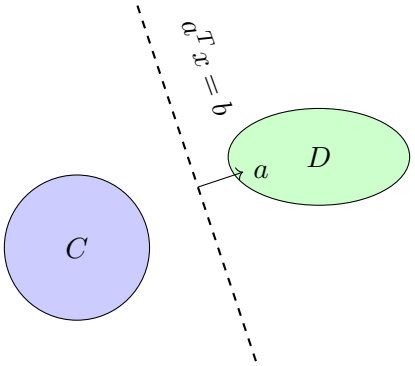


Figure 3: The Separating Hyperplane Theorem. The vector a is normal to the hyperplane separating C and D.

Proof. Let $S = C - D = \{x - u \mid x \in C, u \in D\}$. Since C and D are convex, S is convex. Since $C \cap D = \emptyset$, $0 \notin S$. We use two facts on the projection Π_K onto a nonempty closed convex set K: it exists and is unique (admitted), and $p = \Pi_K(z)$ satisfies

$$\langle y - p, z - p \rangle \leq 0 \quad \forall y \in K\,. \tag{1.1}$$

Indeed, for $y \in K$ and $s \in (0,1]$, $p + s(y-p) \in K$, so $\|z - p - s(y-p)\|^2 \geq \|z-p\|^2$, that is $-2s\langle z-p, y-p\rangle + s^2\|y-p\|^2 \geq 0$; divide by s and let $s \to 0^+$.

Step 1: a unit vector a with $a^T y \leq 0$ for all $y \in S$. Let $K = \mathbf{cl}(S)$, which is closed and convex. If $0 \notin K$, let $p = \Pi_K(0) \neq 0$ and $a = -p/\|p\|$. By (1.1) with $z = 0$, $\langle y - p, -p\rangle \leq 0$,

so $a^T y \leq a^T p = -\|p\| < 0$ for all $y \in S$. If $0 \in K \setminus S$, we use that $\mathbf{int}(\mathbf{cl}S) = \mathbf{int}(S)$ for a convex set $S \subseteq \mathbb{R}^n$ (admitted). Since $0 \notin S$, $0 \notin \mathbf{int}(K)$, so there are points $z_k \notin K$ with $z_k \to 0$. Let $p_k = \Pi_K(z_k) \neq z_k$ and $a_k = (z_k - p_k)/\|z_k - p_k\|$. By (1.1), for $y \in S$, $a_k^T y \leq a_k^T p_k = a_k^T z_k - \|z_k - p_k\| \leq a_k^T z_k \leq \|z_k\|$. The unit vectors a_k have a convergent subsequence, with limit a, $\|a\| = 1$. Passing to the limit gives $a^T y \leq 0$ for all $y \in S$.

Step 2: separation. For $x \in C$ and $u \in D$, $x - u \in S$, so $a^T x \leq a^T u$. Hence $b = \sup_{x \in C} a^T x$ is finite (C and D are nonempty), $a^T x \leq b$ on C and $a^T u \geq b$ on D. □

1.2 Bridging Sets and Functions: The Epigraph

The connection between convex sets and convex functions is made by the *epigraph*, the set of points lying on or above the graph of a function. Throughout, functions may take the value $+\infty$, and $\mathbf{dom}(f) = \{x : f(x) < +\infty\}$; f is convex if $\mathbf{dom}(f)$ is convex and $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ for all $x, y \in \mathbf{dom}(f)$ and $\theta \in [0,1]$.

Definition 1.4 (Epigraph)**.** The epigraph of a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the subset of $\mathbb{R}^{n+1}$ defined by:

$$\mathbf{epi}(f) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq t\} \tag{1.2}$$

Theorem 1.5. *A function f is convex if and only if its epigraph $\mathbf{epi}(f)$ is a convex set.*

Proof. (⇒) Assume f is a convex function. Let (x_1, t_1) and (x_2, t_2) be two arbitrary points in $\mathbf{epi}(f)$. By the definition of the epigraph, this implies:

$$f(x_1) \leq t_1 \quad \text{and} \quad f(x_2) \leq t_2$$

Let $\theta \in [0,1]$. We must show that the convex combination of these points lies in $\mathbf{epi}(f)$. Consider the point:

$$(x_\theta, t_\theta) = \theta(x_1, t_1) + (1-\theta)(x_2, t_2) = (\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2)$$

Evaluate f at the spatial coordinate x_θ. By the definition of a convex function:

$$f(x_\theta) = f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2)$$

Substituting the inequalities $f(x_1) \leq t_1$ and $f(x_2) \leq t_2$:

$$f(x_\theta) \leq \theta t_1 + (1-\theta)t_2 = t_\theta$$

Since $f(x_\theta) \leq t_\theta$, the point $(x_\theta, t_\theta) \in \mathbf{epi}(f)$. Thus, $\mathbf{epi}(f)$ is a convex set.

(⇐) Assume $\mathbf{epi}(f)$ is a convex set. Then $\mathbf{dom}(f)$, the projection of $\mathbf{epi}(f)$ on the first n coordinates, is convex. Let $x_1, x_2 \in \mathbf{dom}(f)$ and $\theta \in [0,1]$. The points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ clearly belong to $\mathbf{epi}(f)$ since $f(x) \leq f(x)$. Because the set is convex, their convex combination must also belong to $\mathbf{epi}(f)$:

$$(\theta x_1 + (1-\theta)x_2, \theta f(x_1) + (1-\theta)f(x_2)) \in \mathbf{epi}(f)$$

By the definition of the epigraph, for a point (u, v) to be in the set, we must have $f(u) \leq v$. Applying this here:

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2)$$

This is exactly the condition for f being a convex function. □

This theorem lets us apply geometric tools, such as the separating hyperplane theorem, to functions. For instance, if g is a subgradient of f at x (Section 1.4), the vector $(g, -1)$ is normal to a hyperplane that supports the epigraph at $(x, f(x))$, as displayed in Figure 4.

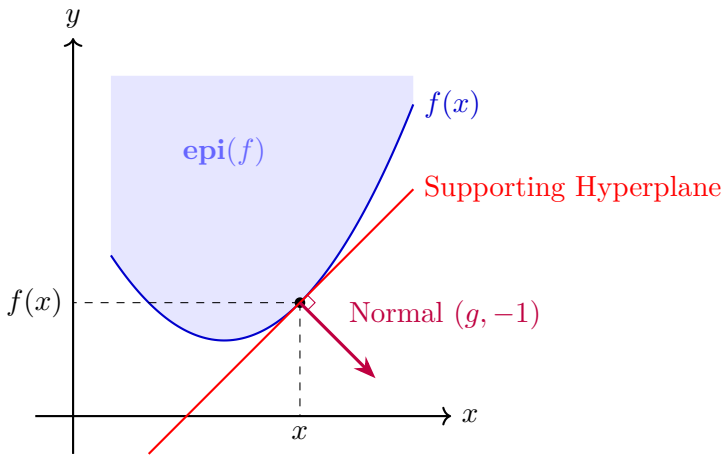


Figure 4: The vector $(g, -1)$ is normal to the supporting hyperplane of the epigraph at $(x, f(x))$; here f is differentiable and $g = f'(x)$.

1.3 Convex Functions: Analysis and Characterization

Recall that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex if its domain is convex and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)\,, \quad x, y \in \mathbf{dom}(f),\ \theta \in [0,1]\,.$$

Geometrically, the chord connecting $(x, f(x))$ and $(y, f(y))$ lies above the graph of the function. A first consequence: *every local minimizer x of a convex function is a global minimizer.* Indeed, for $y \in \mathbf{dom}(f)$ and $t \in (0, 1]$ small enough, $f(x) \leq f(x + t(y-x)) \leq (1-t)f(x) + tf(y)$, hence $f(x) \leq f(y)$.

Theorem 1.6 (First-Order Condition)**.** *Suppose* $\mathbf{dom}(f)$ *is open and f is differentiable on* $\mathbf{dom}(f)$*. Then f is convex if and only if* $\mathbf{dom}(f)$ *is convex and for all $x, y \in \mathbf{dom}(f)$:*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{1.3}$$

Proof. Necessity ($\Rightarrow$): Assume f is convex. By definition, for any $t \in (0, 1]$,

$$f(x + t(y-x)) \leq (1-t)f(x) + tf(y)$$

Rearranging terms, we get:

$$\frac{f(x + t(y-x)) - f(x)}{t} \leq f(y) - f(x)$$

Taking the limit as $t \to 0^+$, the LHS is exactly the definition of the directional derivative in the direction $(y - x)$, which for a differentiable function is $\nabla f(x)^T (y - x)$. Thus, we have the inequality $\nabla f(x)^T (y - x) \leq f(y) - f(x)$.

Sufficiency ($\Leftarrow$): Assume the first-order condition holds. Let $x, y \in \mathbf{dom}(f)$ and set $z = \theta x + (1-\theta)y$. Applying the condition at z toward x and y:

$$f(x) \geq f(z) + \nabla f(z)^T (x - z)$$
$$f(y) \geq f(z) + \nabla f(z)^T (y - z)$$

Multiplying the first inequality by θ and the second by $(1-\theta)$ and summing:

$$\theta f(x) + (1-\theta)f(y) \geq f(z) + \nabla f(z)^T \underbrace{(\theta(x - z) + (1-\theta)(y - z))}_{=0}$$

Thus $f(z) \leq \theta f(x) + (1-\theta)f(y)$, which proves convexity. □

Remark 1.7. The first-order condition says that the first-order Taylor approximation is a *global underestimator* of a convex function. In particular, if $\nabla f(x^\star) = 0$, the inequality becomes $f(y) \geq f(x^\star)$ for all y: any stationary point of a differentiable convex function is a global minimizer.

Theorem 1.8 (Second-Order Condition)**.** *Suppose* $\mathbf{dom}(f)$ *is open and* f *is twice differentiable on* $\mathbf{dom}(f)$*. Then* f *is convex if and only if its domain is convex and its Hessian is positive semidefinite (PSD):*

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbf{dom}(f)$$

Proof. Assume $\mathbf{dom}(f)$ is convex. For $x \in \mathbf{dom}(f)$ and $v \in \mathbb{R}^n$, let $I_{x,v} = \{t : x+tv \in \mathbf{dom}(f)\}$, an open interval containing 0, and $g(t) = f(x+tv)$ on $I_{x,v}$. By definition, f is convex if and only if every such g is convex. A twice differentiable function on an open interval is convex if and only if its second derivative is nonnegative (its derivative is then nondecreasing, which is equivalent to convexity in dimension one). By the chain rule, $g''(t) = v^T \nabla^2 f(x+tv) v$. If $\nabla^2 f \succeq 0$ on $\mathbf{dom}(f)$, every $g'' \geq 0$, so every g is convex and f is convex. Conversely, if f is convex, then $g''(0) = v^T \nabla^2 f(x) v \geq 0$ for all $x \in \mathbf{dom}(f)$ and all v. □

1.4 Subgradients

Definition 1.9 (Subgradient, subdifferential)**.** Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex and $x \in \mathbf{dom}(f)$. A vector $g \in \mathbb{R}^n$ is a *subgradient* of f at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \mathbb{R}^n .$$

The set $\partial f(x)$ of all subgradients at x is the *subdifferential*. It is a closed convex set (an intersection of closed halfspaces). We set $\partial f(x) = \emptyset$ if $x \notin \mathbf{dom}(f)$.

If f is convex and differentiable at $x \in \mathbf{int}(\mathbf{dom} f)$, then $\partial f(x) = \{\nabla f(x)\}$. Indeed, $\nabla f(x) \in \partial f(x)$: the proof of the necessity part of the first-order condition above only uses differentiability at x, and the inequality is trivial for $y \notin \mathbf{dom}(f)$. Conversely, if $g \in \partial f(x)$, then for every v and small $t > 0$, $f(x+tv) - f(x) \geq t\, g^T v$; dividing by t and letting $t \to 0^+$ gives $\nabla f(x)^T v \geq g^T v$ for all v, so $g = \nabla f(x)$ (take v and $-v$).

Proposition 1.10 (Fermat's rule)**.** *Let* f *be convex. Then* $x^\star \in \operatorname{argmin} f$ *if and only if* $0 \in \partial f(x^\star)$.

Proof. $0 \in \partial f(x^\star)$ means $f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$ for all y. □

Proposition 1.11 (Sum rule)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be convex and differentiable, and* $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be convex. For every* $x \in \mathbf{dom}(g)$*,*

$$\partial(f + g)(x) = \nabla f(x) + \partial g(x) .$$

Proof. ($\supseteq$) If $\eta \in \partial g(x)$, add $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ and $g(y) \geq g(x) + \eta^T(y - x)$. ($\subseteq$) Let $\xi \in \partial(f + g)(x)$, $y \in \mathbf{dom}(g)$, $d = y - x$ and $t \in (0, 1]$. By convexity of g, $g(x + td) \leq g(x) + t(g(y) - g(x))$. Hence

$$f(x + td) - f(x) + t(g(y) - g(x)) \geq (f + g)(x + td) - (f + g)(x) \geq t\, \xi^T d .$$

Dividing by t and letting $t \to 0^+$ gives $\nabla f(x)^T d + g(y) - g(x) \geq \xi^T d$, that is $g(y) \geq g(x) + (\xi - \nabla f(x))^T (y - x)$. This also holds if $y \notin \mathbf{dom}(g)$. So $\xi - \nabla f(x) \in \partial g(x)$. □

Definition 1.12 (Dual norm)**.** The dual norm of a norm $\|\cdot\|$ on $\mathbb{R}^n$ is $\|\eta\|_* = \sup_{\|x\| \leq 1} \langle \eta, x \rangle$. By definition, $\langle \eta, x \rangle \leq \|\eta\|_* \|x\|$ for all η, x. The dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$, and the dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$ (the supremum $\max_i |\eta_i|$ is attained at a signed canonical vector $\pm e_i$).

Proposition 1.13 (Subdifferential of a norm)**.** *For every* $x \in \mathbb{R}^n$, $\partial\|\cdot\|(x) = \{\eta : \|\eta\|_* \leq 1,\ \langle \eta, x\rangle = \|x\|\}$. *In particular,*

$$\partial\|\cdot\|_1(x) = \{\eta : \|\eta\|_\infty \leq 1,\ \eta_i = \operatorname{sign}(x_i) \text{ whenever } x_i \neq 0\}\,.$$

Proof. ($\supseteq$) If $\|\eta\|_* \leq 1$ and $\langle \eta, x\rangle = \|x\|$, then $\|x\| + \langle \eta, y - x\rangle = \langle \eta, y\rangle \leq \|y\|$ for all y. ($\subseteq$) Let $\eta \in \partial\|\cdot\|(x)$. With $y = 0$: $0 \geq \|x\| - \langle \eta, x\rangle$. With $y = 2x$: $2\|x\| \geq \|x\| + \langle \eta, x\rangle$. So $\langle \eta, x\rangle = \|x\|$, and then the subgradient inequality reads $\|y\| \geq \langle \eta, y\rangle$ for all y, that is $\|\eta\|_* \leq 1$. For $\|\cdot\|_1$: if $\|\eta\|_\infty \leq 1$, then $\langle \eta, x\rangle = \sum_i \eta_i x_i \leq \sum_i |x_i|$, with equality if and only if $\eta_i x_i = |x_i|$ for all i, that is $\eta_i = \operatorname{sign}(x_i)$ whenever $x_i \neq 0$. □

1.5 Curvature Refined: Smoothness and Strong Convexity

While convexity guarantees that a local minimum is global, the *speed* of optimization algorithms depends on the curvature of the function, bounded by two quadratic approximations.

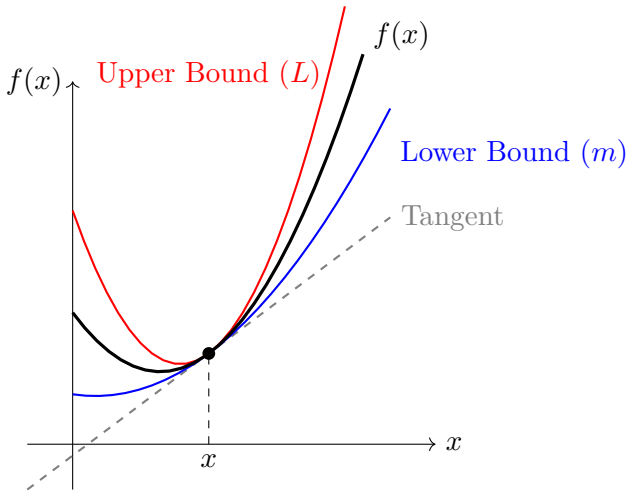


Figure 5: Conditioning: an L-smooth and m-strongly convex function (black) lies between the wide parabola defined by m (blue) and the narrow parabola defined by L (red), both tangent to f at x; this holds at every base point x.

Definition 1.14 (L-Smoothness)**.** A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if its gradient is Lipschitz continuous with constant L: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y. This implies the *quadratic upper bound* (Lemma 1.16 below):

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2\,. \tag{1.4}$$

Definition 1.15 (m-Strong Convexity)**.** Let $m > 0$. A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex if it satisfies the *quadratic lower bound*

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2 \quad \forall x, y\,. \tag{1.5}$$

Equivalently, $f - \frac{m}{2}\|\cdot\|^2$ is convex (apply the first-order condition to it).

If f is twice differentiable, f is m-strongly convex and L-smooth if and only if the eigenvalues of the Hessian lie in $[m, L]$:

$$mI \preceq \nabla^2 f(x) \preceq LI \quad \forall x\,. \tag{1.6}$$

(Apply the second-order condition to $f - \frac{m}{2}\|\cdot\|^2$; and, for $\nabla^2 f \succeq 0$, ∇f is L-Lipschitz if and only if $\|\nabla^2 f(x)\|_{op} \leq L$ for all x.) The ratio $\kappa = L/m \geq 1$ is the *condition number*. If $\kappa \approx 1$,

the level sets are close to spheres. If $\kappa \gg 1$, they are elongated ellipsoids (a narrow valley), and gradient methods are slow (Section 2).

L-smoothness implies the following *descent lemma*, used throughout these notes:

Lemma 1.16 (Descent Lemma)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, then for all x, y:*

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2$$

Proof. By the fundamental theorem of calculus, $f(y) = f(x) + \int_0^1 \nabla f(x + \tau(y - x))^T(y - x)d\tau$. We can rewrite this as $f(x) + \nabla f(x)^T(y - x) + \int_0^1 (\nabla f(x + \tau(y - x)) - \nabla f(x))^T(y - x)d\tau$. Using Cauchy-Schwarz and $\|\nabla f(x + \tau(y - x)) - \nabla f(x)\| \leq L\tau\|y - x\|$:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T(y - x) + \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|\|y - x\|d\tau \\ &\leq f(x) + \nabla f(x)^T(y - x) + \int_0^1 L\tau\|y - x\|^2 d\tau \\ &= f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2 \end{aligned}$$

□

1.6 Relaxing Convexity: The Polyak-Lojasiewicz (PL) Condition

Strong convexity gives linear convergence of gradient descent (Section 2), but it is a restrictive assumption; it fails for non-convex objectives such as those of neural networks. A weaker condition (implied by strong convexity, Theorem 1.19) that still gives linear convergence is the *Polyak-Lojasiewicz (PL) inequality.*

Definition 1.17 (PL Condition)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable with $f^\star = \inf f > -\infty$, and let $\mu > 0$. The function f satisfies the μ-PL condition if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star) \quad \forall x\,. \tag{1.7}$$

Remark 1.18 (Geometry of PL)*.* Unlike strong convexity, the definition of the PL condition *does not involve convexity.* It does not imply a unique minimizer (it allows a valley of global minima), but it implies that every stationary point ($\nabla f(x) = 0$) is a global minimizer, since then $f(x) - f^\star \leq 0$.

Theorem 1.19 (Strong Convexity implies PL)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and m-strongly convex, then $f^\star = \inf f > -\infty$ and f satisfies the Polyak-Lojasiewicz condition with constant $\mu = m$.*

Proof. By the definition of m-strong convexity, for any x, y:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2 \tag{1.8}$$

Since this holds for *all* y, it must hold for the specific y that minimizes the quadratic right-hand side. Let us view the RHS as a function of y, denoted $q(y)$. The minimum of this quadratic occurs where its gradient is zero:

$$\nabla_y q(y) = \nabla f(x) + m(y - x) = 0 \implies y^\star = x - \frac{1}{m}\nabla f(x)$$

Substituting $y^\star$ back into the inequality gives a global lower bound on $f(y)$:

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T \left(-\frac{1}{m}\nabla f(x)\right) + \frac{m}{2}\left\|-\frac{1}{m}\nabla f(x)\right\|^2 \\ &= f(x) - \frac{1}{m}\|\nabla f(x)\|^2 + \frac{1}{2m}\|\nabla f(x)\|^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \end{aligned}$$

Since this inequality holds for all y, we may take the infimum over y:

$$f^\star \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 > -\infty\,.$$

Rearranging terms yields the PL condition:

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

□

Theorem 1.20 (Convergence under PL)**.** *If f is L-smooth and satisfies the μ-PL condition, gradient descent $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ (Section 2.1) converges linearly:*

$$f(x_k) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^\star) \tag{1.9}$$

Proof. By the Descent Lemma (Lemma 1.16) with $y = x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2\,. \tag{1.10}$$

Subtracting $f^\star$ from both sides:

$$f(x_{k+1}) - f^\star \leq f(x_k) - f^\star - \frac{1}{2L}\|\nabla f(x_k)\|^2 \tag{1.11}$$

The μ-PL condition (Definition 1.17) states that $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$. We can rearrange this to bound the squared gradient norm:

$$-\|\nabla f(x_k)\|^2 \leq -2\mu(f(x_k) - f^\star) \tag{1.12}$$

Substituting this into the inequality:

$$f(x_{k+1}) - f^\star \leq f(x_k) - f^\star - \frac{1}{2L}\left(2\mu(f(x_k) - f^\star)\right) \tag{1.13}$$

$$= \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^\star) \tag{1.14}$$

Applying this recurrence relation recursively k times yields the result:

$$f(x_k) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^\star) \tag{1.15}$$

□

1.7 Duality and Optimality

Duality replaces the constraints of the primal problem by penalties weighted by multipliers. Consider the primal problem in standard form:

$$(P): \quad p^\star = \inf_x f_0(x) \quad \text{s.t.} \quad f_i(x) \le 0,\ i = 1, \dots, m, \quad h_j(x) = 0,\ j = 1, \dots, p\,, \tag{1.16}$$

with $f_0, f_i, h_j : \mathbb{R}^n \to \mathbb{R}$. In this subsection, m is the number of inequality constraints (elsewhere it denotes a strong convexity constant). The problem is *convex* when f_0, f_i are convex and the h_j are affine. The first results below do not need convexity.

The Lagrangian and Dual Function

We define the *Lagrangian* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as:

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \nu_j h_j(x) \tag{1.17}$$

The *Lagrange dual function* $g(\lambda, \nu) \in [-\infty, +\infty)$ is the infimum of the Lagrangian over x:

$$g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu) \tag{1.18}$$

Theorem 1.21 (Concavity of the Dual)**.** *The dual function $g(\lambda, \nu)$ is concave, even if the primal problem is not convex.*

Proof. For each x, $(\lambda, \nu) \mapsto \mathcal{L}(x, \lambda, \nu)$ is affine. A pointwise infimum of affine functions is concave: for $\theta \in [0, 1]$, $\inf_x(\theta a_x + (1-\theta) b_x) \ge \theta \inf_x a_x + (1-\theta) \inf_x b_x$. □

Weak and Strong Duality

The dual problem is to maximize the lower bound provided by the dual function:

$$(D): \quad d^\star = \sup_{\lambda \ge 0, \nu} g(\lambda, \nu) \tag{1.19}$$

Theorem 1.22 (Weak Duality)**.** *For any primal feasible x and any (λ, ν) with $\lambda \ge 0$, $g(\lambda, \nu) \le f_0(x)$. Consequently*

$$d^\star \le p^\star\,. \tag{1.20}$$

Proof. Let x be primal feasible and $\lambda \ge 0$.

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x'} \Big(f_0(x') + \sum \lambda_i f_i(x') + \sum \nu_j h_j(x') \Big) \\ &\le f_0(x) + \sum \lambda_i f_i(x) + \sum \nu_j h_j(x) \le f_0(x)\,, \end{aligned}$$

since $h_j(x) = 0$, $f_i(x) \le 0$ and $\lambda_i \ge 0$. Taking the infimum over feasible x gives $g(\lambda, \nu) \le p^\star$, then the supremum over $\lambda \ge 0, \nu$ gives $d^\star \le p^\star$. □

Remark 1.23 (Duality gap as a stopping test). For a primal feasible x and a dual feasible (λ, ν), weak duality gives $f_0(x) - p^\star \le f_0(x) - g(\lambda, \nu)$. The computable gap $f_0(x) - g(\lambda, \nu) \ge 0$ therefore bounds the suboptimality of x and serves as a stopping test. The duality gap of the problem is $p^\star - d^\star \ge 0$.

Strong duality ($d^\star = p^\star$) usually requires convexity and a constraint qualification, such as Slater's condition.

Theorem 1.24 (Slater, admitted)**.** *Assume the problem is convex and there exists x with $f_i(x) < 0$ for all i and $h_j(x) = 0$ for all j (a strictly feasible point; with functions taking values in $\mathbb{R} \cup \{+\infty\}$, one also asks $x \in \mathbf{relint} \bigcap_i \mathbf{dom}(f_i)$). Then $d^\star = p^\star$, and if $d^\star > -\infty$ the dual problem has a solution.*

KKT Conditions: Necessity and Sufficiency

The Karush-Kuhn-Tucker (KKT) conditions characterize primal and dual solutions when strong duality holds.

Theorem 1.25 (KKT Conditions)**.** *Suppose strong duality holds ($p^\star = d^\star$, finite). Let $x^\star$ be primal optimal and $(\lambda^\star, \nu^\star)$ be dual optimal. Then they satisfy:*

1. Stationarity: *$x^\star$ minimizes $x \mapsto \mathcal{L}(x, \lambda^\star, \nu^\star)$; for a convex problem, this is $0 \in \partial_x \mathcal{L}(x^\star, \lambda^\star, \nu^\star)$ (Proposition 1.10);*
2. Primal feasibility: *$f_i(x^\star) \le 0, h_j(x^\star) = 0$;*
3. Dual feasibility: *$\lambda_i^\star \ge 0$;*
4. Complementary slackness: *$\lambda_i^\star f_i(x^\star) = 0$ for all i.*

Conversely, if the problem is convex and $(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ satisfy 1–4, then $\tilde{x}$ is primal optimal, $(\tilde{\lambda}, \tilde{\nu})$ is dual optimal and $p^\star = d^\star$.

Proof of Necessity ($\Rightarrow$). Let $x^\star$ and $(\lambda^\star, \nu^\star)$ be optimal with zero duality gap.

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) \quad \text{(strong duality)} \\ &= \inf_x \mathcal{L}(x, \lambda^\star, \nu^\star) \\ &\le f_0(x^\star) + \sum \lambda_i^\star f_i(x^\star) + \sum \nu_j^\star h_j(x^\star) \quad \text{(Evaluating at } x^\star\text{)} \\ &\le f_0(x^\star) \quad \text{(Since } \lambda^\star \ge 0, f_i \le 0, h_j = 0\text{)} \end{aligned}$$

All inequalities must be equalities. Equality in the first one means that $x^\star$ minimizes $\mathcal{L}(\cdot, \lambda^\star, \nu^\star)$ (stationarity). Equality in the last one gives $\sum \lambda_i^\star f_i(x^\star) = 0$. Since all terms are non-positive, each is zero (complementary slackness). Feasibility holds by assumption. □

Proof of the converse (convex problems). Suppose the problem is convex and $(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ satisfy 1–4. Since $\tilde{\lambda} \ge 0$, f_0, f_i are convex and the h_j are affine, $\mathcal{L}(\cdot, \tilde{\lambda}, \tilde{\nu})$ is convex, and stationarity with Proposition 1.10 shows that $\tilde{x}$ minimizes it. Hence

$$g(\tilde{\lambda}, \tilde{\nu}) = \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) = f_0(\tilde{x}) + \sum \tilde{\lambda}_i f_i(\tilde{x}) + \sum \tilde{\nu}_j h_j(\tilde{x}) = f_0(\tilde{x}) \,,$$

by primal feasibility ($h_j(\tilde{x}) = 0$) and complementary slackness. By weak duality, $f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu}) \le d^\star \le p^\star \le f_0(\tilde{x})$, since $\tilde{x}$ is feasible and $\tilde{\lambda} \ge 0$. All these quantities are equal: $\tilde{x}$ is primal optimal, $(\tilde{\lambda}, \tilde{\nu})$ is dual optimal and $p^\star = d^\star$. □

Example: optimality and duality for the LASSO

Let $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\lambda > 0$, and consider the LASSO

$$P(\beta) = \tfrac{1}{2}\|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \,, \qquad \beta \in \mathbb{R}^d \,. \tag{1.21}$$

We write X_j for the j-th column of X.

Proposition 1.26 (LASSO: optimality, dual, duality gap)**.** *1. $\beta^\star$ minimizes P if and only if $X^T(y - X\beta^\star) \in \lambda\, \partial \|\cdot\|_1(\beta^\star)$, that is, for all j: $|X_j^T(y - X\beta^\star)| \le \lambda$, with $X_j^T(y - X\beta^\star) = \lambda \operatorname{sign}(\beta_j^\star)$ when $\beta_j^\star \ne 0$. In particular, $\beta^\star = 0$ is a solution if and only if $\lambda \ge \|X^T y\|_\infty$.*

2. A dual problem is

$$\max_{p \in \mathbb{R}^n:\ \|X^T p\|_\infty \le 1} D(p) = \lambda \langle p, y \rangle - \frac{\lambda^2}{2}\|p\|_2^2 \,, \tag{1.22}$$

and $P(\beta) \ge D(p)$ for every β and every dual feasible p (weak duality). Strong duality holds, the dual solution $p^\star$ is unique, and $p^\star = (y - X\beta^\star)/\lambda$ for every primal solution $\beta^\star$.

3. *For any β, the point $p(\beta) = (y - X\beta)/\max\left(\lambda, \|X^T(y - X\beta)\|_\infty\right)$ is dual feasible, and the computable gap $P(\beta) - D(p(\beta)) \geq P(\beta) - \min P \geq 0$ is a stopping test.*

Proof. 1. P is the sum of the differentiable convex function $\frac{1}{2}\|y - X\beta\|^2$, with gradient $-X^T(y - X\beta)$, and of $\lambda\|\cdot\|_1$. By Fermat's rule and the sum rule (Propositions 1.10 and 1.11), $\beta^\star$ is optimal if and only if $X^T(y - X\beta^\star) \in \lambda\partial\|\cdot\|_1(\beta^\star)$; Proposition 1.13 gives the coordinatewise form. For $\beta^\star = 0$, $\partial\|\cdot\|_1(0)$ is the unit ℓ_∞ ball, so the condition reads $\|X^T y\|_\infty \leq \lambda$.

2. Write the LASSO as $\min_{\beta,r} \frac{1}{2}\|r\|^2 + \lambda\|\beta\|_1$ subject to $r = y - X\beta$, with multiplier $\lambda p \in \mathbb{R}^n$ for the equality constraint. The Lagrangian is $\frac{1}{2}\|r\|^2 + \lambda\|\beta\|_1 + \lambda\langle p, y - X\beta - r\rangle$. The infimum over r is attained at $r = \lambda p$ and equals $-\frac{\lambda^2}{2}\|p\|^2$. The infimum over β of $\lambda(\|\beta\|_1 - \langle X^T p, \beta\rangle)$ is 0 if $\|X^T p\|_\infty \leq 1$ (since $\langle X^T p, \beta\rangle \leq \|X^T p\|_\infty \|\beta\|_1$) and $-\infty$ otherwise (take $\beta = s\,\mathrm{sign}((X^T p)_j)e_j$ with $|(X^T p)_j| > 1$ and $s \to \infty$). This gives (1.22), and weak duality gives $P(\beta) \geq D(p)$ (the constraint is $r = y - X\beta$, so every β gives a feasible pair). Strong duality: let $\beta^\star$ be a minimizer (it exists since P is continuous and $P(\beta) \geq \lambda\|\beta\|_1$) and $p^\star = (y - X\beta^\star)/\lambda$. By item 1, $\|X^T p^\star\|_\infty \leq 1$ and $\langle X^T p^\star, \beta^\star\rangle = \|\beta^\star\|_1$. Then

$$D(p^\star) = \langle y - X\beta^\star, y\rangle - \tfrac{1}{2}\|y - X\beta^\star\|^2 = \tfrac{1}{2}\|y - X\beta^\star\|^2 + \langle y - X\beta^\star, X\beta^\star\rangle = P(\beta^\star)\,,$$

using $\langle y - X\beta^\star, X\beta^\star\rangle = \lambda\langle X^T p^\star, \beta^\star\rangle = \lambda\|\beta^\star\|_1$. So $p^\star$ is dual optimal and the gap is zero. The dual objective is λ^2-strongly concave, so its maximizer over the closed convex feasible set is unique; since this holds for every primal solution $\beta^\star$, $p^\star = (y - X\beta^\star)/\lambda$ does not depend on the choice of $\beta^\star$.

3. If $c = \max(\lambda, \|X^T(y - X\beta)\|_\infty)$, then $\|X^T p(\beta)\|_\infty = \|X^T(y - X\beta)\|_\infty / c \leq 1$. The inequality is weak duality. □

2 First-Order Methods

Second-order methods (Section 4) cost $O(n^3)$ operations per iteration in dimension n. With millions of parameters, machine learning relies mostly on first-order methods, which only use gradients (or subgradients).

2.1 Gradient Descent (GD)

For an unconstrained problem $\min f(x)$ with f differentiable, gradient descent iterates

$$x_{k+1} = x_k - t_k \nabla f(x_k)\,,$$

where $t_k > 0$ is the step size.

Convergence Analysis

We use the two assumptions of Section 1.5: L-smoothness, $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$, which gives the descent lemma (Lemma 1.16), and m-strong convexity, $f(y) \ge f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2$.

Theorem 2.1 (Convergence Rates)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, and let x_k be the GD iterates with constant step $t = 1/L$.*

1. *If f is convex and has a minimizer $x^\star$, then for $k \ge 1$, $f(x_k) - f^\star \le \frac{L\|x_0 - x^\star\|^2}{2k}$.*
2. *If f is m-strongly convex, with minimizer $x^\star$, then for $k \ge 0$,*

$$\|x_k - x^\star\|^2 \le \Big(1 - \frac{m}{L}\Big)^k \|x_0 - x^\star\|^2 \quad \textit{and} \quad f(x_k) - f^\star \le \Big(1 - \frac{m}{L}\Big)^k (f(x_0) - f^\star)\,.$$

Proof. Write $g_k = \nabla f(x_k)$, so that $x_{k+1} = x_k - g_k/L$. The descent lemma with $y = x_{k+1}$ gives

$$f(x_{k+1}) \le f(x_k) - \frac{1}{L}\|g_k\|^2 + \frac{L}{2}\frac{1}{L^2}\|g_k\|^2 = f(x_k) - \frac{1}{2L}\|g_k\|^2\,. \tag{2.1}$$

In particular, $f(x_k)$ is non-increasing, and (applying (2.1) at any point x in place of x_k, with $f^\star \le f(x - \nabla f(x)/L)$)

$$\|\nabla f(x)\|^2 \le 2L(f(x) - f^\star) \quad \forall x\,. \tag{2.2}$$

Item 1. By convexity, $f(x_k) \le f^\star + \langle g_k, x_k - x^\star\rangle$. With (2.1),

$$f(x_{k+1}) \le f^\star + \langle g_k, x_k - x^\star\rangle - \frac{1}{2L}\|g_k\|^2 = f^\star + \frac{L}{2}\Big(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\Big)\,,$$

where the equality follows from expanding $\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - \frac{2}{L}\langle g_k, x_k - x^\star\rangle + \frac{1}{L^2}\|g_k\|^2$. Summing over $k = 0, \dots, K-1$, the right-hand side telescopes:

$$\sum_{k=1}^{K}(f(x_k) - f^\star) \le \frac{L}{2}\|x_0 - x^\star\|^2\,.$$

Since $f(x_k)$ is non-increasing, the left-hand side is at least $K(f(x_K) - f^\star)$, which gives item 1.

Item 2, function values. By Theorem 1.19, $\|g_k\|^2 \ge 2m(f(x_k) - f^\star)$. With (2.1), $f(x_{k+1}) - f^\star \le (1 - \frac{m}{L})(f(x_k) - f^\star)$; iterate.

Item 2, iterates. Strong convexity between x_k and $x^\star$ gives $f^\star \geq f(x_k) + \langle g_k, x^\star - x_k\rangle + \frac{m}{2}\|x_k - x^\star\|^2$, that is $\langle g_k, x_k - x^\star\rangle \geq f(x_k) - f^\star + \frac{m}{2}\|x_k - x^\star\|^2$. With (2.2),

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star\|^2 - \frac{2}{L}\langle g_k, x_k - x^\star\rangle + \frac{1}{L^2}\|g_k\|^2 \\ &\leq \|x_k - x^\star\|^2 - \frac{2}{L}(f(x_k) - f^\star) - \frac{m}{L}\|x_k - x^\star\|^2 + \frac{2}{L}(f(x_k) - f^\star) \\ &= \Big(1 - \frac{m}{L}\Big)\|x_k - x^\star\|^2 .\end{aligned}$$

Iterating gives the bound on the iterates. □

Remark 2.2 (Conditioning). The factor $1 - m/L = 1 - 1/\kappa$ governs the speed of convergence.

- If $\kappa \approx 1$ (nearly spherical level sets), convergence is fast.
- If $\kappa \gg 1$ (a narrow valley), $1 - 1/\kappa$ is close to 1 and convergence is slow. On a quadratic $f(x) = \frac{1}{2}x^T Q x$ with eigenvalues in $[m, L]$, GD multiplies the coordinate of x_k along an eigenvector of eigenvalue q by $1 - tq$. With $t = 1/L$ the steepest coordinate is set to 0 in one step. With the larger step $t = 2/(L+m)$, the factor is $-(\kappa-1)/(\kappa+1)$ for $q = L$: this coordinate changes sign at each step (zig-zag across the valley), while the coordinate along $q = m$ is multiplied by $(\kappa - 1)/(\kappa + 1)$. This motivates momentum and preconditioning.

Remark 2.3 (Session 2 notebook, experiment 1). The notebook runs GD with step $2/(L + m)$ on $f(x) = \frac{1}{2}x^T Q x$ from $x_0 = (2, 2)$, with $Q = \mathrm{diag}(1, 1)$ and $Q = \mathrm{diag}(10, 1)$. For $\kappa = 1$, the step is 1 and GD converges in one iteration. For $\kappa = 10$, each coordinate is multiplied by $\pm 9/11 \approx \pm 0.818$ at each iteration (zig-zag on the steep coordinate), and 35 iterations are needed to reach $f(x_k) \leq 10^{-6} f(x_0)$ (`figures/fig_td2_gd_conditioning.pdf`).

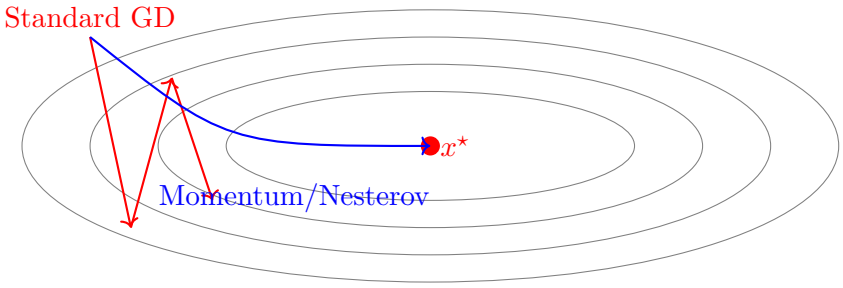


Figure 6: Schematic picture of momentum: on an ill-conditioned problem, GD (red) may oscillate across the narrow valley, while momentum methods (blue) average successive directions. This is an illustration, not a computed trajectory.

2.2 Stochastic Gradient Descent (SGD)

In machine learning, the objective often takes the form of a finite sum over n data points:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) , \tag{2.3}$$

where f_i is the (differentiable) loss associated with the i-th observation. For large n, computing the full gradient $\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x)$ at every iteration is expensive.

Stochastic gradient descent replaces the gradient by the gradient of a single term (or a small batch) at each step:

$$x_{k+1} = x_k - t_k \nabla f_{i_k}(x_k) , \tag{2.4}$$

where the indices $i_0, i_1, \ldots$ are independent and uniformly distributed on $\{1, \ldots, n\}$. In particular i_k is independent of x_k, which depends only on $i_0, \ldots, i_{k-1}$.

Convergence Analysis

The analysis uses that the stochastic gradient is an *unbiased estimator* of the gradient: $\mathbb{E}[\nabla f_{i_k}(x_k) \mid x_k] = \nabla f(x_k)$. Its variance adds noise, and the analysis below gives sublinear rates instead of the linear rate of GD.

Assumption 2.4 (Bounded Variance)**.** There exists a constant σ^2 such that for all x:

$$\mathbb{E}_i[\|\nabla f_i(x) - \nabla f(x)\|^2] \leq \sigma^2 \,, \tag{2.5}$$

where $\mathbb{E}_i$ denotes the expectation with respect to the uniform random index $i \in \{1, \dots, n\}$.

Assumption 2.5 (Bounded second moment along the iterates)**.** There exists $G \geq 0$ such that $\mathbb{E}\|\nabla f_{i_k}(x_k)\|^2 \leq G^2$ for all k.

Since $\mathbb{E}_i\|\nabla f_i(x)\|^2 = \|\nabla f(x)\|^2 + \mathbb{E}_i\|\nabla f_i(x) - \nabla f(x)\|^2$, Assumption 2.5 holds with $G^2 = \sigma^2 + \sup_x \|\nabla f(x)\|^2$ when this supremum is finite, for instance when f is Lipschitz. For an m-strongly convex f on $\mathbb{R}^n$ this supremum is $+\infty$, since $\|\nabla f(x)\| \geq m\|x - x^\star\|$ (by Cauchy-Schwarz and (2.7) below). A standard way to satisfy the assumption is projected SGD, $x_{k+1} = \Pi_{\mathcal{X}}(x_k - t_k \nabla f_{i_k}(x_k))$, on a compact convex set $\mathcal{X}$ containing $x^\star$: then $G^2 = \sigma^2 + \sup_{\mathcal{X}} \|\nabla f\|^2$, and the proofs below are unchanged because $\|\Pi_{\mathcal{X}}(z) - x^\star\| \leq \|z - x^\star\|$ (the projection is non-expansive, by (1.1), and $\Pi_{\mathcal{X}}(x^\star) = x^\star$).

Theorem 2.6 (SGD for Strongly Convex Functions)**.** *Assume f is L-smooth and m-strongly convex, with minimizer $x^\star$, and that Assumption 2.5 holds. With the step size $t_k = \frac{1}{m(k+1)}$, for all $k \geq 1$:*

$$\mathbb{E}\|x_k - x^\star\|^2 \leq \frac{G^2}{m^2 k} \qquad \textit{and} \qquad \mathbb{E}[f(x_k) - f^\star] \leq \frac{LG^2}{2m^2 k} \,. \tag{2.6}$$

Proof. 1. One step. Expanding the square,

$$\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - 2t_k \langle \nabla f_{i_k}(x_k), x_k - x^\star \rangle + t_k^2 \|\nabla f_{i_k}(x_k)\|^2 \,.$$

Taking the conditional expectation given x_k in the middle term (unbiasedness), then the total expectation, and using Assumption 2.5:

$$\mathbb{E}\|x_{k+1} - x^\star\|^2 \leq \mathbb{E}\|x_k - x^\star\|^2 - 2t_k \, \mathbb{E}\langle \nabla f(x_k), x_k - x^\star \rangle + t_k^2 G^2 \,.$$

2. Strong convexity. Adding the strong convexity inequalities between x and $x^\star$ in both directions, and using $\nabla f(x^\star) = 0$:

$$\langle \nabla f(x), x - x^\star \rangle \geq m\|x - x^\star\|^2 \quad \forall x \,. \tag{2.7}$$

With $\delta_k = \mathbb{E}\|x_k - x^\star\|^2$, we get

$$\delta_{k+1} \leq (1 - 2mt_k)\delta_k + t_k^2 G^2 \,.$$

3. Solving the recursion. With $t_k = \frac{1}{m(k+1)}$, $1 - 2mt_k = \frac{k-1}{k+1}$ and

$$\delta_{k+1} \leq \frac{k-1}{k+1}\,\delta_k + \frac{G^2}{m^2(k+1)^2} \,.$$

For $k = 0$ the coefficient is -1, so $\delta_1 \leq -\delta_0 + G^2/m^2 \leq G^2/m^2$. For $k \geq 1$ the coefficient is non-negative; if $\delta_k \leq \frac{G^2}{m^2 k}$, then

$$\delta_{k+1} \leq \frac{G^2}{m^2(k+1)}\Big(\frac{k-1}{k} + \frac{1}{k+1}\Big) = \frac{G^2}{m^2(k+1)}\Big(1 - \frac{1}{k(k+1)}\Big) \leq \frac{G^2}{m^2(k+1)} \,.$$

By induction, $\delta_k \leq \frac{G^2}{m^2 k}$ for all $k \geq 1$.

4. Function values. The descent lemma at $x^\star$, with $\nabla f(x^\star) = 0$, gives $f(x) - f^\star \leq \frac{L}{2}\|x - x^\star\|^2$. Taking expectations gives the second bound. □

Remark 2.7 (Constant step and noise floor). With a constant step $t \leq \frac{1}{2m}$, the recursion of step 2 gives $\delta_k \leq (1-2mt)^k\delta_0 + t^2G^2\sum_{j\geq 0}(1-2mt)^j = (1-2mt)^k\delta_0 + \frac{tG^2}{2m}$. This bound decreases linearly to the level $\frac{tG^2}{2m}$, proportional to t, and does not go to zero: this is the "noise floor" of constant-step SGD. Decreasing steps remove it. The *Robbins-Monro conditions*

$$\sum_{k=0}^{\infty} t_k = \infty, \quad \sum_{k=0}^{\infty} t_k^2 < \infty \tag{2.8}$$

are a classical sufficient (not necessary) condition on the steps for the convergence of SGD under suitable assumptions (admitted) . The step of Theorem 2.6 satisfies them.

Remark 2.8 (Mini-batching). In practice, one often uses a *mini-batch* B_k of b indices drawn independently and uniformly:

$$g_k = \frac{1}{b}\sum_{j\in B_k} \nabla f_j(x_k)\,. \tag{2.9}$$

Then g_k is still unbiased, and its variance $\mathbb{E}\|g_k - \nabla f(x_k)\|^2$ is divided by b (variance of an average of b independent terms), at b times the cost per step. Mini-batches also allow vectorized computations (GPUs).

Theorem 2.9 (SGD for General Convex Functions)**.** *Assume f is convex with a minimizer $x^\star$. Let $R > 0$ with $R \geq \|x_0 - x^\star\|$, and let $G > 0$ be such that Assumption 2.5 holds, with this G, for the iterates of SGD run with any constant step size $t > 0$ (so G does not depend on the step; for instance $G^2 = \sigma^2 + \sup_x \|\nabla f(x)\|^2$ when this supremum is finite, as shown after Assumption 2.5). If we run SGD for T iterations with the constant step size $t = \frac{R}{G\sqrt{T}}$, the* average iterate

$$\bar{x}_T = \frac{1}{T}\sum_{k=0}^{T-1} x_k$$

satisfies

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{RG}{\sqrt{T}}\,. \tag{2.10}$$

Proof. As in step 1 of the previous proof, with constant step t:

$$\mathbb{E}\|x_{k+1} - x^\star\|^2 \leq \mathbb{E}\|x_k - x^\star\|^2 - 2t\,\mathbb{E}\langle \nabla f(x_k), x_k - x^\star\rangle + t^2G^2\,.$$

By convexity, $f^\star \geq f(x_k) + \langle \nabla f(x_k), x^\star - x_k\rangle$, that is $\langle \nabla f(x_k), x_k - x^\star\rangle \geq f(x_k) - f^\star$. Hence

$$2t\,\mathbb{E}[f(x_k) - f^\star] \leq \mathbb{E}\|x_k - x^\star\|^2 - \mathbb{E}\|x_{k+1} - x^\star\|^2 + t^2G^2\,.$$

Summing from $k = 0$ to $T-1$, the distance terms telescope:

$$2t\sum_{k=0}^{T-1}\mathbb{E}[f(x_k) - f^\star] \leq \|x_0 - x^\star\|^2 - \mathbb{E}\|x_T - x^\star\|^2 + Tt^2G^2 \leq R^2 + Tt^2G^2\,.$$

By convexity (Jensen's inequality), $f(\bar{x}_T) \leq \frac{1}{T}\sum_{k=0}^{T-1} f(x_k)$. Dividing by $2tT$:

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{R^2}{2tT} + \frac{tG^2}{2}\,.$$

The step $t = R/(G\sqrt{T})$ minimizes the right-hand side over $t > 0$ and gives $\frac{RG}{2\sqrt{T}} + \frac{RG}{2\sqrt{T}} = \frac{RG}{\sqrt{T}}$. □

Remark 2.10 (Session 2 notebook, experiment 3). On least squares $f(w) = \frac{1}{2N}\|y - Xw\|^2$ with $N = 1000$ and $d = 50$, the notebook compares GD with step $1/L$ and SGD with step $t_k = \min(1/(2L_{\max}),\ 1/(m(k+1)))$, where $L_{\max} = \max_i \|x_i\|^2$. The gaps are plotted against iterations and against epochs (one epoch is N stochastic gradients, the cost of one full gradient). After one epoch, the gap $f - f^\star$ is 2.53 for GD and $3.0 \cdot 10^{-2}$ for SGD; after 20 epochs it is $9.5 \cdot 10^{-11}$ for GD and $2.9 \cdot 10^{-4}$ for SGD (`figures/fig_td2_epochs_iterations.pdf`).

2.3 Variance Reduction (SVRG)

With a constant step, the bounds for SGD stop at a noise floor because of the variance of the gradient estimator. *Variance reduction* builds an estimator whose variance goes to zero when the iterates approach $x^\star$, which allows a constant step size and a linear rate. The *SVRG* (Stochastic Variance Reduced Gradient) method periodically computes a full gradient at a "snapshot" point and uses it as a control variate.

Definition 2.11 (SVRG Estimator)**.** Let $\tilde{x}$ be a snapshot point where the full gradient $\nabla f(\tilde{x})$ is known. The SVRG direction is:

$$\tilde{g}_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla f(\tilde{x})\,. \tag{2.11}$$

Given x_k and $\tilde{x}$, $\mathbb{E}[\tilde{g}_k \mid x_k, \tilde{x}] = \nabla f(x_k)$ (it is unbiased). Its variance is at most $\mathbb{E}_i\|\nabla f_i(x_k) - \nabla f_i(\tilde{x})\|^2 \le L^2\|x_k - \tilde{x}\|^2$ if each ∇f_i is L-Lipschitz (the variance of a random vector is at most its second moment), so it vanishes when x_k and $\tilde{x}$ both approach $x^\star$.

The algorithm runs in epochs $s = 1, 2, \dots$ of length M. At epoch s: set $\tilde{x} = \tilde{x}_{s-1}$, compute $\nabla f(\tilde{x})$, set $x_0 = \tilde{x}$, run $x_{j+1} = x_j - \eta\,\tilde{g}_j$ for $j = 0, \dots, M-1$, and set $\tilde{x}_s = x_j$ for j drawn uniformly in $\{0, \dots, M-1\}$. One epoch costs $n + 2M$ gradient evaluations of the f_i.

Theorem 2.12 (Linear Convergence of SVRG, admitted)**.** *Assume each f_i is convex and L-smooth, and f is m-strongly convex. Let $\eta < \frac{1}{4L}$ and M be large enough so that*

$$\rho = \frac{1}{m\eta(1-2L\eta)M} + \frac{2L\eta}{1-2L\eta} < 1\,.$$

Then $\mathbb{E}[f(\tilde{x}_s) - f^\star] \le \rho^s[f(\tilde{x}_0) - f^\star]$ for all $s \ge 0$.

For instance, $\eta = \frac{1}{10L}$ and $M = 50\kappa$ (with $\kappa = L/m$) give $m\eta(1-2L\eta)M = \frac{1}{10}\cdot\frac{4}{5}\cdot 50 = 4$, hence $\rho = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. With $\eta = \frac{1}{4L}$ the second term of ρ equals 1, so the theorem never applies.

Remark 2.13 (Session 2 notebook, experiment 4). On the least-squares problem of experiment 3, SVRG with $\eta = 1/(4L_{\max})$ and $M = N$ decreases the gap from 20.4 to $1.4 \cdot 10^{-16}$ in 45 epochs (one outer loop costs $N + 2M = 3$ epochs). This step is outside the range of Theorem 2.12 (here $\rho = 2.10$): the linear convergence is observed, not guaranteed. On this problem $L/m = 2.4$ while $L_{\max}/m = 138$, and GD is faster than SVRG per epoch (`figures/fig_td2_sgd_svrg.pdf`).

2.4 The Subgradient Method

When f is convex but not differentiable (e.g., $f(x) = |x|$), we use a *subgradient* $g \in \partial f(x)$ (Section 1.4):

$$f(y) \ge f(x) + g^T(y - x) \quad \forall y\,.$$

Algorithm: $x_{k+1} = x_k - t_k g_k$, where $g_k \in \partial f(x_k)$.

Theorem 2.14 (Subgradient Method at a Fixed Horizon)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and G-Lipschitz (then $\|g\| \le G$ for all subgradients), with a minimizer $x^\star$, and let $R \ge \|x_0 - x^\star\|$. Fix a horizon $k \ge 0$ and use the constant step $t_i = t = \frac{R}{G\sqrt{k+1}}$ for $i = 0, \dots, k$. Then*

$$\min_{i=0,\dots,k}\{f(x_i) - f^\star\} \le \frac{RG}{\sqrt{k+1}}\,.$$

Proof. If $g \in \partial f(x)$ and $g \ne 0$, take $y = x + g/\|g\|$: $G \ge f(y) - f(x) \ge \|g\|$, so $\|g\| \le G$. For each i,

$$\|x_{i+1} - x^\star\|^2 = \|x_i - x^\star\|^2 - 2t_i g_i^T(x_i - x^\star) + t_i^2\|g_i\|^2\,.$$

The subgradient inequality at x_i, evaluated at $x^\star$, gives $g_i^T(x_i - x^\star) \geq f(x_i) - f^\star$, hence

$$\|x_{i+1} - x^\star\|^2 \leq \|x_i - x^\star\|^2 - 2t_i(f(x_i) - f^\star) + t_i^2 G^2 \,.$$

Summing over $i = 0, \ldots, k$ and using $\|x_{k+1} - x^\star\|^2 \geq 0$:

$$2\sum_{i=0}^{k} t_i(f(x_i) - f^\star) \leq \|x_0 - x^\star\|^2 + G^2 \sum_{i=0}^{k} t_i^2 \,.$$

With $f_{best}^{(k)} = \min_{i \leq k} f(x_i)$, the left-hand side is at least $2(f_{best}^{(k)} - f^\star)\sum_i t_i$, so

$$f_{best}^{(k)} - f^\star \leq \frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2\sum_{i=0}^{k} t_i} \,.$$

With $t_i = R/(G\sqrt{k+1})$, the right-hand side is $\frac{R^2+R^2}{2(k+1)R/(G\sqrt{k+1})} = \frac{RG}{\sqrt{k+1}}$. □

Remark 2.15 (Session 2 notebook, experiment 2). On $f(x) = |x|$ from $x_0 = 2.05$ ($G = 1$), the notebook compares the constant step 0.2 and the decreasing step $1/\sqrt{k+1}$, together with the bound $(R^2 + G^2 \sum t_i^2)/(2\sum t_i)$. With the constant step, $f_{best}^{(k)}$ stays at 0.05 from $k = 10$ on. With the decreasing step, $f_{best}^{(k)} = 0.131$, 0.068, 0.040 at $k = 10, 25, 50$, below the bound, about 0.34 at $k = 50$ (`figures/fig_td2_subgradient.pdf`).

2.5 Adaptive Methods: Adam and AdamW

SGD uses the same step size t_k for all parameters. In deep learning, parameters often differ in scale (e.g., word embeddings vs. dense layers). *Adaptive gradient methods* adjust the step size of each coordinate from the history of its gradients. The most used method of this class is *Adam* (Adaptive Moment Estimation), which combines momentum (first moment) and RMSProp (second moment) . No convergence theorem for Adam is given in these notes. In the algorithm below, g_k^2 and the division are taken coordinatewise.

Algorithm 1 AdamW (Adam with Decoupled Weight Decay)

Require: Stepsize α, decay rates $\beta_1, \beta_2 \in [0, 1)$, ϵ, weight decay λ

1: Initialize moment vectors $m_0 = 0, v_0 = 0$
2: **for** $k = 1, \ldots, T$ **do**
3: $\quad g_k = \nabla f_{i_k}(x_{k-1})$ (Stochastic Gradient)
4: $\quad m_k = \beta_1 m_{k-1} + (1 - \beta_1) g_k$ (Momentum)
5: $\quad v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k^2$ (Squared Gradients)
6: $\quad \hat{m}_k = m_k/(1 - \beta_1^k), \quad \hat{v}_k = v_k/(1 - \beta_2^k)$ (Bias Correction)
7: $\quad$ *Update:*

$$x_k = x_{k-1} - \alpha \left(\frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon} + \lambda x_{k-1} \right)$$

8: **end for**

Remark 2.16 (Why AdamW?). Adam with ℓ_2 regularization adds the term λx_{k-1} to the gradient g_k. In adaptive methods, this term is then divided by $\sqrt{\hat{v}_k} + \epsilon$, coordinate by coordinate, so the effective regularization differs across coordinates. *AdamW* decouples the weight decay: the term λx_{k-1} is added directly in the update, outside the adaptive scaling .

Remark 2.17 (Session 2 notebook, experiment 5). On the Rosenbrock function ($a = 1$, $b = 100$) from $(-1.5, 2.5)$, with exact gradients, the notebook compares GD (step 0.001), GD with heavy-ball momentum (0.9, step 0.001) and Adam (step 0.1). After 500 iterations, $f(x_k)$ equals 4.88 for GD, $6.2 \cdot 10^{-3}$ for heavy ball and $1.1 \cdot 10^{-2}$ for Adam (`figures/fig_td2_rosenbrock.pdf`). On this example, Adam does not beat heavy ball.

2.6 Coordinate Descent (CD)

Cyclic coordinate descent minimizes the objective with respect to one variable at a time, keeping the others fixed; for $j = 1, \dots, n$:

$$x_j^{k+1} = \operatorname*{argmin}_z f(x_1^{k+1}, \dots, x_{j-1}^{k+1}, z, x_{j+1}^k, \dots, x_n^k)\,.$$

Convergence. For a general non-smooth convex function, CD can stop at a non-optimal point: for $f(x_1, x_2) = |x_1 - x_2| + 0.1(x_1 + x_2)$, CD started at $(0, 0)$ never moves ($t \mapsto |t| \pm 0.1t$ is minimized at $t = 0$), while $f(-s, -s) = -0.2s < 0$ for $s > 0$. For $F = f + \sum_j g_j(x_j)$ with f convex and differentiable and closed convex g_j (such as the ℓ_1 norm), the limit points of CD are minimizers if F is bounded below, $F(x^0) < +\infty$, and there is $\mu > 0$ such that, for every j and every value of the other coordinates, $z \mapsto f(x_1, \dots, x_{j-1}, z, x_{j+1}, \dots, x_n)$ is μ-strongly convex (admitted). For the LASSO (1.21), this holds when every column X_j is nonzero, with $\mu = \min_j \|X_j\|^2$; since $\lambda > 0$, the iterates then stay in a bounded set, so limit points exist. No rate is proved in these notes.

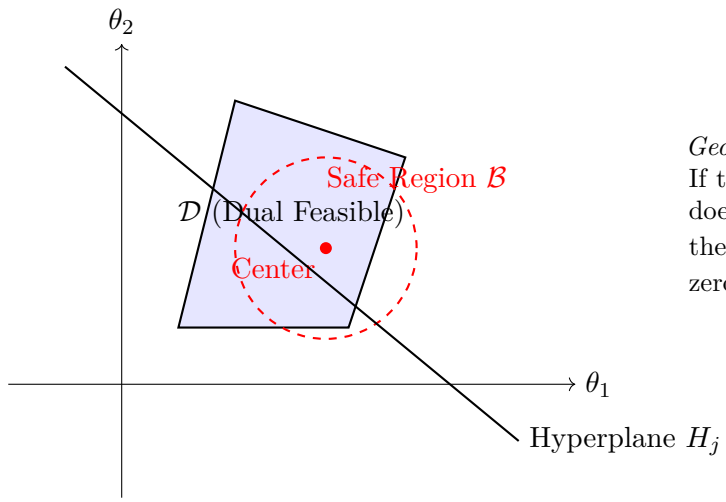
CD for the LASSO (1.21). With $r_j = y - \sum_{l \neq j} X_l \beta_l$, the coordinate problem is $\min_b \frac{1}{2}\|r_j - X_j b\|^2 + \lambda |b|$. Its optimality condition $X_j^T r_j - \|X_j\|^2 b \in \lambda\, \partial|\cdot|(b)$ gives

$$\beta_j \leftarrow \frac{S_\lambda(X_j^T r_j)}{\|X_j\|^2}\,, \qquad S_\lambda(u) = \operatorname{sign}(u) \max(|u| - \lambda, 0)\,,$$

the soft-thresholding operator (Example 3.2). If the residual $y - X\beta$ is stored and updated after each coordinate, one update costs $O(n_{\text{obs}})$, where n_{obs} is the number of rows of X, and one sweep over the d coordinates costs $O(n_{\text{obs}} d)$, the cost of one gradient. The `glmnet` package uses this method .

Remark 2.18 (Session 3 notebook, experiment 2). On a LASSO with $n_{\text{obs}} = 100$, $d = 2000$ and $\lambda = \lambda_{\max}/10$, where $\lambda_{\max} = \|X^T y\|_\infty$, the notebook compares CD (scikit-learn, random order) and ISTA. CD reaches a gap $F - F^\star$ of 2.3 after 84 epochs and $2.5 \cdot 10^{-11}$ after 1062 epochs; ISTA is at $1.3 \cdot 10^3$ after 100 iterations and 8.7 after 1000 iterations (`figures/fig_td3_cd_ista.pdf`). The figure also shows the gap against time; timings depend on the machine.

2.7 Screening Rules (SAFE)



Geometric Screening:
If the Safe Region $\mathcal{B}$ (containing $\theta^\star$) does not intersect the hyperplane H_j, then the j-th coefficient is guaranteed zero.

Figure 7: Safe screening (schematic): if a ball $\mathcal{B}$ known to contain the dual solution $\theta^\star$ does not meet the hyperplanes $H_j = \{\theta : |X_j^T \theta| = \lambda\}$, then $\beta_j^\star = 0$.

Screening uses duality to find coefficients of the LASSO (1.21) that are zero at the optimum, before or during the optimization, and to remove the corresponding columns of X. We use the dual variable $\theta = \lambda p$ of Proposition 1.26: the dual problem reads $\max_\theta D(\theta) = \frac{1}{2}\|y\|^2 - \frac{1}{2}\|y-\theta\|^2$ subject to $\|X^T\theta\|_\infty \le \lambda$, and its solution is $\theta^\star = y - X\beta^\star$.

Proposition 2.19 (A safe screening rule)**.** *Let $\beta \in \mathbb{R}^d$ and let θ be dual feasible ($\|X^T\theta\|_\infty \le \lambda$). If*

$$|X_j^T\theta| + \|X_j\|\sqrt{2(P(\beta) - D(\theta))} < \lambda\,,$$

then $\beta_j^\star = 0$ for every solution $\beta^\star$ of the LASSO.

Proof. D is 1-strongly concave and $\theta^\star$ maximizes it over the convex feasible set, so $\langle \nabla D(\theta^\star), \theta - \theta^\star\rangle \le 0$ and $D(\theta) \le D(\theta^\star) + \langle \nabla D(\theta^\star), \theta - \theta^\star\rangle - \frac{1}{2}\|\theta-\theta^\star\|^2 \le D(\theta^\star) - \frac{1}{2}\|\theta-\theta^\star\|^2$. By weak duality $D(\theta^\star) \le P(\beta)$, so $\|\theta - \theta^\star\| \le \sqrt{2(P(\beta) - D(\theta))}$. Then $|X_j^T\theta^\star| \le |X_j^T\theta| + \|X_j\|\|\theta - \theta^\star\| < \lambda$. By Proposition 1.26 (item 1), $\beta_j^\star \neq 0$ would force $|X_j^T(y - X\beta^\star)| = |X_j^T\theta^\star| = \lambda$. Hence $\beta_j^\star = 0$. □

The rule is safe: it never removes a variable of the support. As (β, θ) approach optimality, the gap goes to zero and the rule can discard many variables .

2.8 Frank-Wolfe (Conditional Gradient)

In some applications (e.g., matrix completion), the feasible set $\mathcal{C}$ is such that the projection onto $\mathcal{C}$ is expensive: for the nuclear norm ball, it requires a full SVD, $O(n^3)$ operations for an $n \times n$ matrix.

The *Frank-Wolfe* algorithm avoids projections. It relies on a *Linear Minimization Oracle (LMO)*, which is often much cheaper. Starting from $x_0 \in \mathcal{C}$, for $k = 0, 1, \ldots$:

1. compute the gradient $\nabla f(x_k)$;
2. *LMO step:* find a point of $\mathcal{C}$ that minimizes the linear approximation:
$$s_k \in \operatorname*{argmin}_{s\in\mathcal{C}} \langle s, \nabla f(x_k)\rangle\,; \tag{2.12}$$
3. update with step size $\gamma_k = \frac{2}{k+2}$:
$$x_{k+1} = (1-\gamma_k)x_k + \gamma_k s_k\,. \tag{2.13}$$

The iterates stay in $\mathcal{C}$ (convex combinations of points of $\mathcal{C}$).

Theorem 2.20 (Frank-Wolfe rate)**.** *Let $\mathcal{C}$ be a nonempty compact convex set of diameter $D = \max_{x,y\in\mathcal{C}} \|x - y\|$, and f be convex and L-smooth, with minimum $f^\star$ over $\mathcal{C}$. Then for all $k \ge 1$,*

$$f(x_k) - f^\star \le \frac{2LD^2}{k+2}\,.$$

The Frank-Wolfe gap $\mathrm{g}_k = \langle \nabla f(x_k), x_k - s_k\rangle$ *is computable and satisfies $\mathrm{g}_k \ge f(x_k) - f^\star$.*

Proof. Let $x^\star \in \operatorname{argmin}_{\mathcal{C}} f$ and $h_k = f(x_k) - f^\star$. By convexity and the definition of s_k, $f^\star \ge f(x_k) + \langle \nabla f(x_k), x^\star - x_k\rangle \ge f(x_k) + \langle \nabla f(x_k), s_k - x_k\rangle$, which is the statement on g_k. The descent lemma with $x_{k+1} - x_k = \gamma_k(s_k - x_k)$ and $\|s_k - x_k\| \le D$ gives

$$f(x_{k+1}) \le f(x_k) + \gamma_k\langle \nabla f(x_k), s_k - x_k\rangle + \frac{L}{2}\gamma_k^2 D^2 \le f(x_k) - \gamma_k h_k + \frac{LD^2}{2}\gamma_k^2\,,$$

that is $h_{k+1} \le (1-\gamma_k)h_k + \frac{LD^2}{2}\gamma_k^2$. For $k = 0$, $\gamma_0 = 1$ gives $h_1 \le \frac{LD^2}{2} \le \frac{2LD^2}{3}$. If $h_k \le \frac{2LD^2}{k+2}$ for some $k \ge 1$, then

$$h_{k+1} \le \frac{k}{k+2}\cdot\frac{2LD^2}{k+2} + \frac{2LD^2}{(k+2)^2} = \frac{2LD^2(k+1)}{(k+2)^2} \le \frac{2LD^2}{k+3}\,,$$

since $(k+1)(k+3) \le (k+2)^2$. Induction concludes. □

The constant LD^2 can be replaced by a smaller curvature constant $C_f \leq LD^2$.

Example 2.21 (LMO on the ℓ_1 ball)**.** For $\mathcal{C} = \{x : \|x\|_1 \leq \tau\}$, a solution of the LMO is $s = -\tau \operatorname{sign}(\partial_j f(x_k))\, e_j$ with $j \in \operatorname{argmax}_i |\partial_i f(x_k)|$, since $\min_{\|s\|_1 \leq \tau} \langle s, v \rangle = -\tau \|v\|_\infty$ (dual norms, Section 1.4). Each iterate x_k is a convex combination of x_0 and at most k such points, so it has at most k more non-zero entries than x_0.

Example 2.22 (Matrix completion)**.** For the nuclear norm ball $\{\|X\|_* \leq \tau\}$, a solution of the LMO is $-\tau\, u_1 v_1^T$, where (u_1, v_1) is a pair of leading singular vectors of $\nabla f(X_k)$. It can be approximated by the power method, at a cost per iteration proportional to the number of nonzero entries for sparse gradients, compared to a full SVD for the projection.

3 Proximal Algorithms and Operator Splitting Methods

High-dimensional statistics often leads to *composite optimization* problems of the form:

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) + g(x)\,, \tag{3.1}$$

where f is convex and L-smooth, and g is convex but *non-smooth* (e.g., the ℓ_1 norm or the indicator function of a convex set). The subgradient method applied to F has the rate $O(1/\sqrt{k})$ (Theorem 2.14). Proximal algorithms use the smoothness of f to recover the $O(1/k)$ rate of gradient descent (Theorem 3.7). Operator splitting methods, which build on the duality of Section 1.7, split the problem into simpler subproblems solved one after the other.

3.1 The Proximal Operator: A Majorization Perspective

To minimize F, we cannot use the gradient of g (which may not exist). Since f is L-smooth, the descent lemma gives, for a step size $t \le 1/L$, the upper bound

$$f(y) \le \hat{f}_t(y, x_k) = f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2t}\|y - x_k\|_2^2 \quad \forall y\,. \tag{3.2}$$

We build a surrogate of F by replacing f with this upper bound and keeping g exact:

$$y_{k+1} = \operatorname*{argmin}_y \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2t}\|y - x_k\|_2^2 + g(y)\right\}. \tag{3.3}$$

Ignoring constant terms independent of y and completing the square, this is equivalent to:

$$y_{k+1} = \operatorname*{argmin}_y \left\{ g(y) + \frac{1}{2t}\|y - (x_k - t\nabla f(x_k))\|_2^2\right\}. \tag{3.4}$$

So the next iterate is obtained by applying the *proximal operator* of g to a gradient descent step on f.

Definition 3.1 (Proximal Operator)**.** Let $g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a proper (not identically $+\infty$), closed (lower semi-continuous) and convex function. The proximal operator of g with parameter $t > 0$, evaluated at $v \in \mathbb{R}^n$, is defined as:

$$\mathbf{prox}_{g,t}(v) = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ g(x) + \frac{1}{2t}\|x - v\|_2^2\right\}. \tag{3.5}$$

The function minimized is lower semi-continuous and strongly convex, so the minimizer exists and is unique (admitted for existence; uniqueness follows from strict convexity). The point $\mathbf{prox}_{g,t}(v)$ is a compromise between minimizing g and staying close to v.

By Fermat's rule and the sum rule (Propositions 1.10 and 1.11), applied to the differentiable function $\frac{1}{2t}\|\cdot - v\|^2$ and to g,

$$p = \mathbf{prox}_{g,t}(v) \iff \frac{v - p}{t} \in \partial g(p)\,. \tag{3.6}$$

Example 3.2 (Soft-thresholding and projection)**.**

1. For $g = \lambda\|\cdot\|_1$, the problem separates over coordinates and $\mathbf{prox}_{g,t}(v)_i = S_{t\lambda}(v_i) = \operatorname{sign}(v_i)\max(|v_i| - t\lambda, 0)$ (*soft-thresholding*). Indeed, by (3.6) and Proposition 1.13, p_i must satisfy $v_i - p_i \in t\lambda\,\partial|\cdot|(p_i)$: if $p_i > 0$ then $p_i = v_i - t\lambda$, which requires $v_i > t\lambda$; if $p_i < 0$ then $p_i = v_i + t\lambda$, which requires $v_i < -t\lambda$; and $p_i = 0$ requires $|v_i| \le t\lambda$. These three cases cover all v_i, and the prox is unique.
2. For the indicator $g = \iota_{\mathcal{C}}$ of a nonempty closed convex set (0 on $\mathcal{C}$, $+\infty$ outside), $\mathbf{prox}_{g,t}(v) = \operatorname{argmin}_{x\in\mathcal{C}}\|x - v\| = \Pi_{\mathcal{C}}(v)$ for every $t > 0$.

Definition 3.3 (Fenchel-Legendre Conjugate)**.** The convex conjugate of a function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, denoted g^*, is defined by:

$$g^*(y) = \sup_{z \in \mathbb{R}^n} \{\langle z, y\rangle - g(z)\} \tag{3.7}$$

The conjugate g^* is convex and lower semi-continuous (a supremum of affine functions of y), even if g is not. It is used in duality theory, in primal-dual algorithms (such as Chambolle-Pock, Section 3.10), and in the Moreau identity linking $\mathbf{prox}_g$ to $\mathbf{prox}_{g^*}$. In the PGD course, the star $\star$ denotes the conjugate (and the adjoint); in these notes the conjugate is g^*, the adjoint is K^T, and $x^\star$, $f^\star$ denote a minimizer and the optimal value.

Proposition 3.4 (Fenchel-Young inequality and three conjugates)**.** *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex with $\mathbf{dom}(g) \neq \emptyset$.*

1. *For all x, η: $g(x) + g^*(\eta) \geq \langle x, \eta\rangle$. For $x \in \mathbf{dom}(g)$, equality holds if and only if $\eta \in \partial g(x)$.*
2. *For a norm, $\|\cdot\|^*(\eta) = 0$ if $\|\eta\|_* \leq 1$ and $+\infty$ otherwise: the conjugate of a norm is the indicator of the dual unit ball.*
3. *For $h(x) = \frac{1}{2}\|x - y\|_2^2$, $h^*(\eta) = \frac{1}{2}\|\eta\|_2^2 + \langle \eta, y\rangle$.*

Proof. 1. The inequality is the definition of $g^*(\eta)$ as a supremum, evaluated at $z = x$. For $x \in \mathbf{dom}(g)$, equality means that the supremum is attained at x: $\langle z, \eta\rangle - g(z) \leq \langle x, \eta\rangle - g(x)$ for all z, that is $g(z) \geq g(x) + \langle \eta, z - x\rangle$ for all z, that is $\eta \in \partial g(x)$.

2. If $\|\eta\|_* \leq 1$, then $\langle \eta, x\rangle - \|x\| \leq (\|\eta\|_* - 1)\|x\| \leq 0$, with equality at $x = 0$: the supremum is 0. If $\|\eta\|_* > 1$, there is x_0 with $\|x_0\| \leq 1$ and $\langle \eta, x_0\rangle > 1 \geq \|x_0\|$; then $\langle \eta, sx_0\rangle - \|sx_0\| = s(\langle \eta, x_0\rangle - \|x_0\|) \to +\infty$ as $s \to +\infty$.

3. The function $x \mapsto \langle \eta, x\rangle - \frac{1}{2}\|x - y\|^2$ is concave and differentiable, with gradient $\eta - (x - y)$, which vanishes at $x = y + \eta$. The maximum value is $\langle \eta, y + \eta\rangle - \frac{1}{2}\|\eta\|^2 = \frac{1}{2}\|\eta\|^2 + \langle \eta, y\rangle$. □

3.2 ISTA: Iterative Shrinkage-Thresholding a.k.a. Forward-Backward (FB)

ISTA alternates between a gradient descent step on f and a proximal step on g:

$$x_{k+1} = \mathbf{prox}_{g, t_k}\left(x_k - t_k \nabla f(x_k)\right) \tag{3.8}$$

Theorem 3.5 (Sufficient Descent Lemma)**.** *Let f be L-smooth, g be proper, closed and convex, $x \in \mathbf{dom}(g)$, and let the step size satisfy $t \leq 1/L$. Then the ISTA update $x^+ = \mathbf{prox}_{g,t}(x - t\nabla f(x))$ satisfies*

$$F(x^+) \leq F(x) - \frac{1}{2t}\|x^+ - x\|_2^2 \tag{3.9}$$

Proof. Let $u = x - t\nabla f(x)$. By definition of the proximal update, $x^+ = \mathbf{prox}_{g,t}(u) = \operatorname{argmin}_z \left\{g(z) + \frac{1}{2t}\|z - u\|_2^2\right\}$.

Because f is L-smooth, we apply the Descent Lemma (Lemma 1.16), which guarantees:

$$f(x^+) \leq f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{L}{2}\|x^+ - x\|_2^2 . \tag{3.10}$$

Next, we use the optimality condition (3.6) of the proximal step: there exists a subgradient $\xi \in \partial g(x^+)$ satisfying:

$$\xi + \frac{1}{t}(x^+ - u) = 0 \implies \xi = -\nabla f(x) - \frac{1}{t}(x^+ - x) .$$

By the subgradient inequality $g(z) \geq g(x^+) + \langle \xi, z - x^+ \rangle$ applied at $z = x$:

$$\begin{aligned} g(x) &\geq g(x^+) + \langle -\nabla f(x) - \frac{1}{t}(x^+ - x), x - x^+ \rangle \\ &= g(x^+) + \langle \nabla f(x), x^+ - x \rangle + \frac{1}{t}\|x^+ - x\|_2^2 . \end{aligned}$$

Rearranging yields:

$$g(x^+) \leq g(x) - \langle \nabla f(x), x^+ - x \rangle - \frac{1}{t}\|x^+ - x\|_2^2 . \tag{3.11}$$

Summing inequalities (3.10) and (3.11):

$$F(x^+) = f(x^+) + g(x^+) \leq F(x) + \left(\frac{L}{2} - \frac{1}{t}\right)\|x^+ - x\|_2^2 .$$

Since $t \leq 1/L$, we have $\frac{1}{t} \geq L$, so $\frac{L}{2} - \frac{1}{t} \leq -\frac{1}{2t}$. Therefore:

$$F(x^+) \leq F(x) - \frac{1}{2t}\|x^+ - x\|_2^2 ,$$

concluding the proof. In particular $F(x_k)$ is non-increasing along ISTA with constant step $t \leq 1/L$. □

Lemma 3.6 (Key inequality of the proximal gradient step)**.** *Let f be convex and L-smooth, g be proper, closed and convex, $F = f + g$, and $t \leq 1/L$. For $y \in \mathbb{R}^n$, let $x^+ = \mathbf{prox}_{g,t}(y - t\nabla f(y))$. Then for every $z \in \mathbb{R}^n$:*

$$F(x^+) - F(z) \leq \frac{1}{2t}\left(\|y - z\|_2^2 - \|x^+ - z\|_2^2\right) . \tag{3.12}$$

Proof. By the descent lemma and $L \leq 1/t$:

$$F(x^+) \leq f(y) + \langle \nabla f(y), x^+ - y \rangle + \frac{1}{2t}\|x^+ - y\|_2^2 + g(x^+) . \tag{3.13}$$

By (3.6), $\xi = \frac{1}{t}(y - x^+) - \nabla f(y) \in \partial g(x^+)$. The subgradient inequality for g and the first-order condition for the convex function f give, for every z:

$$F(z) \geq f(y) + \langle \nabla f(y), z - y \rangle + g(x^+) + \langle \xi, z - x^+ \rangle . \tag{3.14}$$

Subtracting (3.14) from (3.13), and using $\nabla f(y) + \xi = \frac{1}{t}(y - x^+)$:

$$F(x^+) - F(z) \leq \langle \nabla f(y) + \xi, x^+ - z \rangle + \frac{1}{2t}\|x^+ - y\|_2^2 = \frac{1}{t}\langle y - x^+, x^+ - z \rangle + \frac{1}{2t}\|y - x^+\|_2^2 .$$

With $a = y - x^+$ and $b = x^+ - z$, $2\langle a, b \rangle + \|a\|^2 = \|a + b\|^2 - \|b\|^2 = \|y - z\|^2 - \|x^+ - z\|^2$, which gives (3.12). If $z \notin \mathbf{dom}(g)$, (3.12) is trivial. □

Theorem 3.7 (Rate of ISTA)**.** *Let f be convex and L-smooth, g be proper, closed and convex, and let $x^\star$ be a minimizer of $F = f + g$. ISTA with constant step $t \leq 1/L$ and $x_0 \in \mathbf{dom}(g)$ satisfies, for $k \geq 1$,*

$$F(x_k) - F(x^\star) \leq \frac{\|x_0 - x^\star\|_2^2}{2tk} .$$

Proof. Apply Lemma 3.6 with $y = x_{i-1}$ (so that $x^+ = x_i$) and $z = x^\star$: $F(x_i) - F(x^\star) \leq \frac{1}{2t}(\|x_{i-1} - x^\star\|^2 - \|x_i - x^\star\|^2)$. Summing over $i = 1, \ldots, k$, the right-hand side telescopes and is at most $\frac{1}{2t}\|x_0 - x^\star\|^2$. Since $F(x_i)$ is non-increasing (Theorem 3.5), the left-hand side is at least $k(F(x_k) - F(x^\star))$. □

3.3 FISTA (Fast ISTA) and Nesterov Acceleration

ISTA converges at the rate $O(1/k)$. A lower bound of Nesterov states that, for smooth convex problems and as long as k is small compared to the dimension, no first-order method can guarantee a better rate than order $L\|x_0 - x^\star\|^2/k^2$ (admitted) . Nesterov's momentum attains this order. *FISTA* applies it to ISTA: the proximal gradient step is taken at an extrapolated point y_k instead of the previous iterate . With $g = 0$, FISTA is Nesterov's accelerated gradient method.

Algorithm 2 FISTA

Require: x_0, step size $t \le 1/L$, $s_1 = 1, y_1 = x_0$

1: **for** $k = 1, 2, \dots$ **do**

2: *Gradient Step:* $z_k = y_k - t\nabla f(y_k)$

3: *Proximal Step:* $x_k = \mathbf{prox}_{g,t}(z_k)$

4: *Momentum Update:*

$$s_{k+1} = \frac{1 + \sqrt{1 + 4s_k^2}}{2}, \quad \beta_k = \frac{s_k - 1}{s_{k+1}}$$

5: *Extrapolation:* $y_{k+1} = x_k + \beta_k(x_k - x_{k-1})$

6: **end for**

Remark 3.8 (Inertia). The term $\beta_k(x_k - x_{k-1})$ adds "inertia" to the trajectory: the step keeps part of the previous displacement. Unlike ISTA, FISTA is not a descent method: $F(x_k)$ may increase at some iterations (Section 5).

Remark 3.9 (Admissible inertia scaling). The proof below only uses $s_1 = 1$, $s_k \ge \frac{k+1}{2}$ and $s_{k+1}^2 - s_{k+1} \le s_k^2$. The sequences $s_k = \frac{k+1}{2}$ and $\beta_k = \frac{s_k - 1}{s_{k+1}} = \frac{k-1}{k+2}$ satisfy them, since $\left(\frac{k+2}{2}\right)^2 - \frac{k+2}{2} = \frac{k^2+2k}{4} \le \frac{(k+1)^2}{4}$; they give the same rate.

Theorem 3.10 (Convergence Rate of FISTA)**.** *Let f be convex and L-smooth, and let g be proper, closed, and convex. Let $F = f + g$ and let $x^\star$ be a minimizer of F. The FISTA algorithm with step size $t = 1/L$ satisfies, for $k \ge 1$:*

$$F(x_k) - F(x^\star) \le \frac{2L\|x_0 - x^\star\|_2^2}{(k+1)^2}\,. \tag{3.15}$$

By the lower bound recalled above, this order $1/k^2$ cannot be improved in general by a first-order method.

Proof. Write $F^\star = F(x^\star)$, $\Delta_k = F(x_k) - F^\star \ge 0$, and, with x_0 as in the algorithm,

$$w_k = x_{k-1} + s_k(x_k - x_{k-1})\,, \qquad \mathcal{E}_k = s_k^2\Delta_k + \frac{1}{2t}\|w_k - x^\star\|_2^2\,, \qquad k \ge 1\,.$$

Step 1: two identities. Since $y_{k+1} = x_k + \frac{s_k - 1}{s_{k+1}}(x_k - x_{k-1})$,

$$s_{k+1}y_{k+1} - (s_{k+1} - 1)x_k = x_k + (s_k - 1)(x_k - x_{k-1}) = w_k\,, \qquad s_{k+1}x_{k+1} - (s_{k+1} - 1)x_k = w_{k+1}\,.$$

Step 2: one-step decrease. Apply Lemma 3.6 at $y = y_{k+1}$ (so that $x^+ = x_{k+1}$) with $z = (1 - s_{k+1}^{-1})x_k + s_{k+1}^{-1}x^\star$. By convexity of F (and $s_{k+1} \ge 1$), $F(z) \le (1 - s_{k+1}^{-1})F(x_k) + s_{k+1}^{-1}F^\star$, so

$$\Delta_{k+1} - (1 - s_{k+1}^{-1})\Delta_k \le F(x_{k+1}) - F(z) \le \frac{1}{2t}\left(\|y_{k+1} - z\|^2 - \|x_{k+1} - z\|^2\right)\,.$$

Multiply by s_{k+1}^2. On the right, $s_{k+1}(y_{k+1} - z) = w_k - x^\star$ and $s_{k+1}(x_{k+1} - z) = w_{k+1} - x^\star$ by Step 1. On the left, $s_{k+1}^2 - s_{k+1} \le s_k^2$ (with equality for the FISTA sequence, since s_{k+1} is the positive root of $s^2 - s - s_k^2 = 0$) and $\Delta_k \ge 0$. Hence

$$s_{k+1}^2 \Delta_{k+1} - s_k^2 \Delta_k \le \frac{1}{2t}\left(\|w_k - x^\star\|^2 - \|w_{k+1} - x^\star\|^2\right), \quad \text{that is} \quad \mathcal{E}_{k+1} \le \mathcal{E}_k .$$

Step 3: initialization. Since $s_1 = 1$, $w_1 = x_1$, and Lemma 3.6 at $y = y_1 = x_0$ with $z = x^\star$ gives $\Delta_1 \le \frac{1}{2t}(\|x_0 - x^\star\|^2 - \|x_1 - x^\star\|^2)$. Hence $\mathcal{E}_1 \le \frac{1}{2t}\|x_0 - x^\star\|^2$.

Step 4: rate. By induction, $s_k \ge \frac{k+1}{2}$: it holds for $k = 1$, and $s_{k+1} = \frac{1+\sqrt{1+4s_k^2}}{2} \ge \frac{1+2s_k}{2} \ge \frac{k+2}{2}$. Therefore

$$\Delta_k \le \frac{\mathcal{E}_k}{s_k^2} \le \frac{\mathcal{E}_1}{s_k^2} \le \frac{4}{(k+1)^2} \cdot \frac{\|x_0 - x^\star\|^2}{2t} = \frac{2L\|x_0 - x^\star\|_2^2}{(k+1)^2} \quad \text{for } t = \frac{1}{L} .$$

□

Remark 3.11 (Session 3 notebook, experiment 1). On a LASSO with $n = 100$ observations, $d = 500$ and $\lambda = 1$, with step $1/L$, the gap $F(x_k) - F^\star$ after 300 iterations is 8.41 for ISTA (bound of Theorem 3.7: 41.7) and $6.3 \cdot 10^{-5}$ for FISTA (bound of Theorem 3.10: 0.553). FISTA is not monotone: F increases at 61 of its 300 iterations, never for ISTA (`figures/fig_td3_ista_fista.pdf`).

3.4 The Proximal Point Algorithm

Many splitting methods derive from the proximal point algorithm (PPA). For a proper, closed, convex function Φ, PPA minimizes Φ by repeatedly applying its proximal operator:

$$x_{k+1} = \mathbf{prox}_{\Phi, t_k}(x_k) = \operatorname*{argmin}_x \left\{\Phi(x) + \frac{1}{2t_k}\|x - x_k\|_2^2\right\} \tag{3.16}$$

If $\operatorname{argmin}\Phi \neq \emptyset$, the iterates converge to a minimizer of Φ for any step sizes $t_k \ge t_{\min} > 0$ (admitted) . Each iteration decreases Φ, since $\Phi(x_{k+1}) + \frac{1}{2t_k}\|x_{k+1} - x_k\|^2 \le \Phi(x_k)$. However, PPA requires the proximal operator of the full function Φ, which is often as hard to compute as the original problem (for instance for $\Phi = f + g$). It serves as a model for the splitting methods below.

Remark 3.12 (Notation Convention for Splitting Methods). In the subsections that follow, the functions f and g denote the two components of the composite objective $F = f + g$. For the convergence theory of Douglas-Rachford Splitting, we assume that f is the strongly convex and smooth component, while g is proper, closed, and convex (but not necessarily smooth). This is consistent with the ISTA/FISTA convention where f is the smooth part and g is the non-smooth regularizer.

3.5 Monotone Operators and the Optimization Connection

The splitting methods of this section (DRS, ADMM, Chambolle-Pock) can be studied with *monotone operator theory*, which connects optimization, fixed-point theory and variational inequalities.

Definition 3.13 (Monotone and Maximal Monotone Operator)**.** A set-valued operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *monotone* if

$$\langle u - v, x - y\rangle \ge 0 \quad \text{for all } (x,u),(y,v) \in \operatorname{gra}(A),$$

where $\operatorname{gra}(A) = \{(x,u) : u \in A(x)\}$ is the graph of A. It is *maximal monotone* if its graph is not properly contained in the graph of any other monotone operator.

From optimization to monotone inclusion. For f convex and differentiable on $\mathbb{R}^n$ and g convex, Fermat's rule and the sum rule (Propositions 1.10 and 1.11) give the optimality condition for $\min_x F(x) = f(x) + g(x)$:

$$0 \in \partial F(x) = \nabla f(x) + \partial g(x),$$

which is a *monotone inclusion* problem: find x such that $0 \in (A + B)(x)$, with $A = \nabla f$ and $B = \partial g$. The subdifferential of a convex function is monotone: adding the subgradient inequalities $g(y) \geq g(x) + \langle u, y - x\rangle$ and $g(x) \geq g(y) + \langle v, x - y\rangle$ gives $\langle u - v, x - y\rangle \geq 0$. If g is proper, closed and convex, ∂g is maximal monotone (Rockafellar's theorem, admitted) .

Two operators built from A link the monotone inclusion framework to the proximal algorithms studied below.

Definition 3.14 (Resolvent)**.** Let $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximal monotone operator. The *resolvent* of A with parameter $t > 0$ is the operator

$$J_{tA} = (\mathrm{Id} + tA)^{-1},$$

that is, $x = J_{tA}(v)$ if and only if $v \in x + tA(x)$. It is single-valued: if $v = x + ta = y + tb$ with $a \in A(x)$, $b \in A(y)$, then $\|x - y\|^2 = -t\langle a - b, x - y\rangle \leq 0$. It is defined on all of $\mathbb{R}^n$ (Minty's theorem, admitted) . For a proper, closed, convex function g and $A = \partial g$, the resolvent coincides with the proximal operator, by (3.6):

$$J_{t\partial g}(v) = \mathbf{prox}_{g,t}(v) = \operatorname*{argmin}_x \left\{ g(x) + \frac{1}{2t}\|x - v\|^2 \right\}.$$

Definition 3.15 (Reflected Resolvent (Reflected Proximal Operator))**.** The *reflected resolvent* of a maximal monotone operator A is defined as

$$R_{tA} = 2J_{tA} - \mathrm{Id}\,.$$

When $A = \partial g$, this becomes the reflected proximal operator $R_{g,t} = 2\mathbf{prox}_{g,t} - \mathrm{Id}$.

Proposition 3.16 (Properties of the Resolvent and Reflected Resolvent)**.** *Let A be maximal monotone with resolvent J_{tA} and reflected resolvent R_{tA}. Then:*

1. *J_{tA} is* firmly non-expansive*: for all $u, v \in \mathbb{R}^n$,*

$$\|J_{tA}(u) - J_{tA}(v)\|^2 + \|(\mathrm{Id} - J_{tA})(u) - (\mathrm{Id} - J_{tA})(v)\|^2 \leq \|u - v\|^2 .$$

2. *R_{tA} is* non-expansive*: $\|R_{tA}(u) - R_{tA}(v)\| \leq \|u - v\|$ for all u, v.*
3. *$x^\star$ is a zero of A (i.e., $0 \in A(x^\star)$) if and only if $x^\star$ is a* fixed point *of J_{tA}.*

Proof. (1) Firm non-expansiveness of J_{tA}. Let $x = J_{tA}(u)$ and $y = J_{tA}(v)$. By definition of the resolvent, $u \in x + tA(x)$ and $v \in y + tA(y)$, so there exist $a \in A(x)$ and $b \in A(y)$ with $u = x + ta$ and $v = y + tb$. The monotonicity of A gives

$$\langle a - b, x - y\rangle \geq 0 .$$

Now compute:

$$\begin{aligned} \|u - v\|^2 &= \|(x - y) + t(a - b)\|^2 \\ &= \|x - y\|^2 + 2t\langle a - b, x - y\rangle + t^2\|a - b\|^2 . \end{aligned}$$

Since $\langle a - b, x - y\rangle \geq 0$, all three terms are non-negative, so $\|x - y\|^2 \leq \|u - v\|^2$, which shows that J_{tA} is non-expansive. For the firm non-expansiveness identity, observe that $(\mathrm{Id} - J_{tA})(u) = u - x = ta$ and $(\mathrm{Id} - J_{tA})(v) = v - y = tb$, so

$$\|J_{tA}(u) - J_{tA}(v)\|^2 + \|(\mathrm{Id} - J_{tA})(u) - (\mathrm{Id} - J_{tA})(v)\|^2 = \|x - y\|^2 + t^2\|a - b\|^2 .$$

From the expansion above, $\|u-v\|^2 = \|x-y\|^2+2t\langle a-b, x-y\rangle+t^2\|a-b\|^2$. Since $\langle a-b, x-y\rangle \geq 0$, we obtain

$$\|x-y\|^2 + t^2\|a-b\|^2 \leq \|u-v\|^2 ,$$

which is precisely the firm non-expansiveness inequality.

(2) Non-expansiveness of R_{tA}. Write $R_{tA} = 2J_{tA} - \mathrm{Id}$. For any u, v:

$$\begin{aligned}\|R_{tA}(u) - R_{tA}(v)\|^2 &= \|2(x-y) - (u-v)\|^2 \\ &= 4\|x-y\|^2 - 4\langle x-y, u-v\rangle + \|u-v\|^2 .\end{aligned}$$

Using $(u-v) = (x-y) + t(a-b)$, we get $\langle x-y, u-v\rangle = \|x-y\|^2 + t\langle x-y, a-b\rangle$, hence

$$\begin{aligned}\|R_{tA}(u) - R_{tA}(v)\|^2 &= 4\|x-y\|^2 - 4\|x-y\|^2 - 4t\langle x-y, a-b\rangle + \|u-v\|^2 \\ &= \|u-v\|^2 - 4t\langle a-b, x-y\rangle .\end{aligned}$$

By monotonicity, $\langle a-b, x-y\rangle \geq 0$, so $\|R_{tA}(u) - R_{tA}(v)\|^2 \leq \|u-v\|^2$.

(3) Fixed-point characterization. If $0 \in A(x^\star)$, then $x^\star \in x^\star + tA(x^\star)$, so $x^\star = (\mathrm{Id} + tA)^{-1}(x^\star) = J_{tA}(x^\star)$. Conversely, if $J_{tA}(x^\star) = x^\star$, then $x^\star \in (\mathrm{Id} + tA)(x^\star)$, which means $x^\star = x^\star + ta$ for some $a \in A(x^\star)$, giving $a = 0$. □

Remark 3.17 (The Reflected Resolvent for Smooth Functions). When g is convex and differentiable and $A = \nabla g$, the resolvent $J_{tA}(u) = x$ satisfies $u = x + t\nabla g(x)$. The reflected resolvent then reads:

$$R_{g,t}(u) = 2x - u = 2x - (x + t\nabla g(x)) = x - t\nabla g(x) ,$$

which is a *gradient descent step* starting from x. So, when f is smooth, Douglas-Rachford splitting below combines a gradient step for f (inside $R_{f,t}$) and a proximal step for g.

Several objects of this course are monotone operators:

- The *subdifferential* ∂g of a proper, closed, convex function g is maximal monotone (Rockafellar's theorem, admitted above).
- The *gradient* ∇f of a convex differentiable function is monotone. If f is also L-smooth, ∇f is $(1/L)$-*co-coercive*: $\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$ (Baillon-Haddad theorem, admitted) .
- The *proximal operator* $\mathbf{prox}_{g,t}$ is the resolvent $J_{t\partial g}$ (Definition 3.14), hence firmly non-expansive. Its reflection $R_{g,t} = 2\mathbf{prox}_{g,t} - \mathrm{Id}$ (Definition 3.15) is non-expansive; this property is used for Douglas-Rachford splitting.
- For the saddle-point problem $\min_x \max_y \mathcal{L}(x,y) = \langle Kx, y\rangle + f(x) - g^*(y)$ of Chambolle-Pock (Section 3.10), the operator $T(x,y) = (\partial f(x) + K^T y,\ \partial g^*(y) - Kx)$ is monotone: it is the sum of the monotone operator $(x,y) \mapsto \partial f(x) \times \partial g^*(y)$ and of a skew-symmetric linear map, for which $\langle (K^T(y-y'), -K(x-x')), (x-x', y-y')\rangle = 0$.

Douglas-Rachford splitting finds a zero of $A + B$ using the resolvents J_{tA} and J_{tB} separately, and ADMM does the same on the dual problem (Theorem 3.24).

3.6 Douglas-Rachford Splitting

When the objective is a sum of two proper, closed, convex functions, $\min_x f(x)+g(x)$, computing $\mathbf{prox}_{f+g,t}$ is in general as hard as the problem itself. Douglas-Rachford Splitting (DRS) uses the proximal operators of f and g separately.

The DRS iteration is defined by the following sequence:

$$x_{k+1} = \mathbf{prox}_{f,t}(z_k) \tag{3.17}$$

$$y_{k+1} = \mathbf{prox}_{g,t}(2x_{k+1} - z_k) \tag{3.18}$$

$$z_{k+1} = z_k + y_{k+1} - x_{k+1} \tag{3.19}$$

Neither f nor g needs to be smooth: DRS only uses their proximal operators.

3.7 ADMM: Alternating Direction Method of Multipliers

For problems coupled by linear constraints, $\min_{x,z} f(x) + g(z)$ subject to $Ax + Bz = c$, with f, g proper, closed and convex, ADMM uses the *augmented Lagrangian*.

Definition 3.18 (Augmented Lagrangian)**.** Given a penalty parameter $\rho > 0$, the augmented Lagrangian of the linearly constrained problem is:

$$\mathcal{L}_\rho(x, z, u) = f(x) + g(z) + \langle u, Ax + Bz - c\rangle + \frac{\rho}{2}\|Ax + Bz - c\|_2^2 . \tag{3.20}$$

ADMM minimizes the augmented Lagrangian in x, then in z (one pass of block coordinate descent), then takes a gradient ascent step of size ρ on the dual variable u:

$$x^{k+1} \in \operatorname*{argmin}_x \mathcal{L}_\rho(x, z^k, u^k) \tag{3.21}$$

$$z^{k+1} \in \operatorname*{argmin}_z \mathcal{L}_\rho(x^{k+1}, z, u^k) \tag{3.22}$$

$$u^{k+1} = u^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \tag{3.23}$$

We assume that these minimizers exist. ADMM is Douglas-Rachford splitting applied to the dual problem (Theorem 3.24). Since x and z are updated in separate steps, ADMM is used for distributed optimization.

Example 3.19 (ADMM for the LASSO)**.** Write the LASSO with data X, y as $\min_{x,z} \frac{1}{2}\|y - Xx\|^2 + \lambda\|z\|_1$ subject to $x - z = 0$: here $f(x) = \frac{1}{2}\|y - Xx\|^2$, $g = \lambda\|\cdot\|_1$, $A = I$, $B = -I$ and $c = 0$. The three steps are explicit:

$$x^{k+1} = (X^TX + \rho I)^{-1}(X^Ty + \rho z^k - u^k)\,, \qquad z^{k+1} = S_{\lambda/\rho}(x^{k+1} + u^k/\rho)\,,$$
$$u^{k+1} = u^k + \rho(x^{k+1} - z^{k+1})\,.$$

The x-step sets the gradient $X^T(Xx - y) + u^k + \rho(x - z^k)$ to zero. Up to a constant, the z-step minimizes $\lambda\|z\|_1 + \frac{\rho}{2}\|z - x^{k+1} - u^k/\rho\|^2$, whose solution is the soft-thresholding of Example 3.2 with threshold λ/ρ.

Remark 3.20 (Session 3 notebook, experiment 3). On a LASSO with 50 observations, 100 features and $\lambda = 0.5$, the gap $F(z^k) - F^\star$ after 200 iterations is $9.1 \cdot 10^{-2}$ for $\rho = 0.1$, $3.9 \cdot 10^{-5}$ for $\rho = 1$ and $1.3 \cdot 10^{-15}$ for $\rho = 10$; the primal residual $\|x^k - z^k\|$ is then $5.3 \cdot 10^{-2}$, $1.8 \cdot 10^{-3}$ and $1.3 \cdot 10^{-8}$ (`figures/fig_td3_admm.pdf`). The best ρ depends on the problem; this example gives no general rule.

3.8 Linear Convergence of Splitting Methods

For general proper, closed, convex f and g, DRS satisfies the sublinear bound $\|z_{k+1} - z_k\|^2 \le \|z_0 - z^\star\|^2/(k+1)$ on its fixed-point residual (Proposition 3.23 below). Under strong convexity and smoothness of one function, it converges linearly. The following theorem gives contraction factors for (generalized) Douglas-Rachford splitting when f is strongly convex and smooth.

Theorem 3.21 (Linear Convergence of Douglas-Rachford Splitting)**.** *Consider the composite optimization problem $\min_v f(v) + g(v)$, where f and g are proper, closed, and convex. Assume further that f is m-strongly convex and L-smooth. Let $\kappa = L/m$ denote the condition number of f. The generalized Douglas-Rachford iteration is defined as:*

$$v_{k+1} = T_\alpha v_k\,, \qquad T_\alpha = (1 - \alpha)\mathrm{Id} + \alpha R_{g,t}R_{f,t}\,, \tag{3.24}$$

where $t > 0$ is the proximal step size, $\alpha > 0$ is the relaxation parameter, and

$$R_{f,t} = 2\mathbf{prox}_{f,t} - \mathrm{Id}$$

is the reflected proximal operator. Let $t = \frac{1}{\sqrt{mL}}$.

1. *For $\alpha = 1$ (Peaceman-Rachford splitting), T_1 has a unique fixed point $v^\star$ and*
$$\|v_{k+1} - v^\star\| \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)\|v_k - v^\star\|\,. \tag{3.25}$$

2. *For the standard Douglas-Rachford splitting, $\alpha = \frac{1}{2}$, $T_{1/2}$ has the same unique fixed point $v^\star$ and*
$$\|v_{k+1} - v^\star\| \leq \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}\|v_k - v^\star\|\,. \tag{3.26}$$

3. *In both cases, $x^\star = \mathbf{prox}_{f,t}(v^\star)$ is the unique minimizer of $f+g$, and $x_{k+1} = \mathbf{prox}_{f,t}(v_k)$ satisfies $\|x_{k+1} - x^\star\| \leq \|v_k - v^\star\|$.*

These factors are upper bounds obtained from the proof below; their tightness is not proved here .

Proof. Step 1: $R_{g,t}$ is non-expansive. Since g is proper, closed and convex, $\mathbf{prox}_{g,t} = J_{t\partial g}$ and $R_{g,t} = 2\mathbf{prox}_{g,t} - \mathrm{Id}$ is non-expansive (Proposition 3.16). Hence any Lipschitz constant δ of $R_{f,t}$ is a Lipschitz constant of $R_{g,t}R_{f,t}$.

Step 2: $R_{f,t}$ is a contraction. Let $u, v \in \mathbb{R}^n$, $x = \mathbf{prox}_{f,t}(u)$ and $y = \mathbf{prox}_{f,t}(v)$. Since f is differentiable, (3.6) gives $u = x + t\nabla f(x)$ and $v = y + t\nabla f(y)$, so that
$$R_{f,t}(u) = 2x - u = x - t\nabla f(x)\,. \tag{3.27}$$
Let $\Delta x = x - y$ and $\Delta_\nabla = \nabla f(x) - \nabla f(y)$. Then
$$\|R_{f,t}(u) - R_{f,t}(v)\|^2 = \|\Delta x - t\Delta_\nabla\|^2 = \|\Delta x\|^2 - 2t\langle \Delta x, \Delta_\nabla\rangle + t^2\|\Delta_\nabla\|^2\,, \tag{3.28}$$
$$\|u - v\|^2 = \|\Delta x + t\Delta_\nabla\|^2 = \|\Delta x\|^2 + 2t\langle \Delta x, \Delta_\nabla\rangle + t^2\|\Delta_\nabla\|^2\,. \tag{3.29}$$
Because f is m-strongly convex and L-smooth, its gradient satisfies (admitted)
$$\langle \Delta x, \Delta_\nabla\rangle \geq \frac{mL}{m+L}\|\Delta x\|^2 + \frac{1}{m+L}\|\Delta_\nabla\|^2\,. \tag{3.30}$$
For $\delta \in [0,1)$, (3.28) and (3.29) show that $\|R_{f,t}(u) - R_{f,t}(v)\|^2 \leq \delta^2\|u-v\|^2$ is equivalent to
$$0 \leq (\delta^2-1)\|\Delta x\|^2 + 2t(\delta^2+1)\langle \Delta x, \Delta_\nabla\rangle + t^2(\delta^2-1)\|\Delta_\nabla\|^2\,.$$
Since $\delta^2 + 1 > 0$, by (3.30) the right-hand side is at least
$$\Big[(\delta^2-1) + 2t(\delta^2+1)\frac{mL}{m+L}\Big]\|\Delta x\|^2 + \Big[t^2(\delta^2-1) + \frac{2t(\delta^2+1)}{m+L}\Big]\|\Delta_\nabla\|^2\,,$$
which is non-negative as soon as both brackets are non-negative, that is
$$\frac{1-\delta^2}{1+\delta^2} \leq \frac{2tmL}{m+L} \quad \text{and} \quad \frac{1-\delta^2}{1+\delta^2} \leq \frac{2}{t(m+L)}\,. \tag{3.31}$$
The minimum of the two right-hand sides is largest when they are equal, that is $t^2 = 1/(mL)$: this is the choice $t = 1/\sqrt{mL}$. Both right-hand sides are then equal to
$$r = \frac{2\sqrt{mL}}{m+L} = \frac{2\sqrt{\kappa}}{\kappa+1} \in (0,1]\,. \tag{3.32}$$
The value of δ with $\frac{1-\delta^2}{1+\delta^2} = r$ is $\delta^2 = \frac{1-r}{1+r} = \frac{\kappa+1-2\sqrt{\kappa}}{\kappa+1+2\sqrt{\kappa}}$, that is
$$\delta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\,. \tag{3.33}$$

So $R_{f,t}$, and therefore $T_1 = R_{g,t}R_{f,t}$, is δ-Lipschitz.

Step 3: the two rates. For $\alpha = 1/2$, by the triangle inequality,

$$\|T_{1/2}(u)-T_{1/2}(v)\| \le \frac{1}{2}\|u-v\|+\frac{1}{2}\|T_1(u)-T_1(v)\| \le \frac{1+\delta}{2}\|u-v\|\,, \qquad \frac{1+\delta}{2} = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}\,. \tag{3.34}$$

Both factors are < 1. By the Banach fixed-point theorem, T_1 and $T_{1/2}$ each have a unique fixed point; since $T_{1/2}v = v$ if and only if $T_1 v = v$, it is the same point $v^\star$. Then $\|v_{k+1} - v^\star\| = \|T_\alpha v_k - T_\alpha v^\star\|$ gives items 1 and 2.

Step 4: the minimizer. Let $x^\star = \mathbf{prox}_{f,t}(v^\star)$ and $y = \mathbf{prox}_{g,t}(2x^\star - v^\star)$. Then $T_1 v^\star = 2y - 2x^\star + v^\star = v^\star$ gives $y = x^\star$. By (3.6), $v^\star - x^\star \in t\,\partial f(x^\star)$ and $(2x^\star - v^\star) - x^\star \in t\,\partial g(x^\star)$. Adding, $0 \in \partial f(x^\star) + \partial g(x^\star) \subseteq \partial(f+g)(x^\star)$ (add the two subgradient inequalities), so $x^\star$ minimizes $f+g$ (Proposition 1.10); it is unique since $f+g$ is strongly convex. Finally, $\mathbf{prox}_{f,t}$ is non-expansive (Proposition 3.16), so $\|x_{k+1} - x^\star\| \le \|v_k - v^\star\|$. □

Remark 3.22 (From the Generalized Douglas-Rachford Iteration to the Standard DRS Updates). The generalized Douglas-Rachford iteration in Theorem 3.21 is

$$v_{k+1} = (1-\alpha)v_k + \alpha\, R_{g,t}\, R_{f,t}(v_k)\,.$$

To connect this to the three-line DRS recursion, set $\alpha = 1/2$ and expand the reflected resolvents $R_{f,t} = 2\mathbf{prox}_{f,t} - \mathrm{Id}$ and $R_{g,t} = 2\mathbf{prox}_{g,t} - \mathrm{Id}$ step by step.

Step 1 (Prox of f): Define $x_{k+1} = \mathbf{prox}_{f,t}(v_k)$. Then

$$R_{f,t}(v_k) = 2x_{k+1} - v_k\,.$$

Step 2 (Prox of g): Define $y_{k+1} = \mathbf{prox}_{g,t}(2x_{k+1} - v_k)$. Then

$$R_{g,t}(R_{f,t}(v_k)) = 2y_{k+1} - (2x_{k+1} - v_k) = 2y_{k+1} - 2x_{k+1} + v_k\,.$$

Step 3 (Shadow update): Setting $\alpha = 1/2$ in the generalized iteration:

$$\begin{aligned} v_{k+1} &= \tfrac{1}{2}\, v_k + \tfrac{1}{2}\, R_{g,t}R_{f,t}(v_k) \\ &= \tfrac{1}{2}\, v_k + \tfrac{1}{2}(2y_{k+1} - 2x_{k+1} + v_k) \\ &= v_k + y_{k+1} - x_{k+1}\,. \end{aligned}$$

Collecting these three lines recovers exactly the standard DRS iteration:

$$\begin{aligned} x_{k+1} &= \mathbf{prox}_{f,t}(v_k)\,, \\ y_{k+1} &= \mathbf{prox}_{g,t}(2x_{k+1} - v_k)\,, \\ v_{k+1} &= v_k + y_{k+1} - x_{k+1}\,. \end{aligned}$$

Thus the variable v_k of the generalized iteration is the "shadow" sequence z_k of DRS. The case $\alpha = 1$ (Peaceman-Rachford) replaces the averaging with a full step: $v_{k+1} = R_{g,t}R_{f,t}(v_k) = 2y_{k+1} - 2x_{k+1} + v_k$, which doubles the correction and gives the smaller contraction factor $\delta = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ of Theorem 3.21. Without strong convexity, convergence of Peaceman-Rachford is not guaranteed: the bounds of this section do not apply when neither f nor g is strongly convex.

Proposition 3.23 (Sublinear rate of DRS)**.** *Let f and g be proper, closed and convex, $t > 0$, and $T = \frac{1}{2}(\mathrm{Id} + R_{g,t}R_{f,t})$, so that DRS reads $z_{k+1} = Tz_k$. Assume T has a fixed point $z^\star$. Then $x^\star = \mathbf{prox}_{f,t}(z^\star)$ minimizes $f+g$, and for all $k \ge 0$,*

$$\|z_{k+1} - z_k\|^2 = \|y_{k+1} - x_{k+1}\|^2 \le \frac{\|z_0 - z^\star\|^2}{k+1}\,.$$

Proof. Step 4 of the previous proof did not use strong convexity: it shows that $x^\star$ minimizes $f+g$. The operator $N = R_{g,t}R_{f,t}$ is non-expansive (Proposition 3.16). For u, v, let $a = u - v$ and $b = Nu - Nv$; then

$$\|Tu - Tv\|^2 + \|(u - Tu) - (v - Tv)\|^2 = \tfrac{1}{4}\|a+b\|^2 + \tfrac{1}{4}\|a-b\|^2 = \tfrac{1}{2}\|a\|^2 + \tfrac{1}{2}\|b\|^2 \le \|u - v\|^2 .$$

With $u = z_i$ and $v = z^\star$: $\|z_{i+1} - z^\star\|^2 + \|z_i - z_{i+1}\|^2 \le \|z_i - z^\star\|^2$. Summing over $i = 0, \dots, k$ gives $\sum_{i=0}^{k} \|z_i - z_{i+1}\|^2 \le \|z_0 - z^\star\|^2$. The same inequality shows that T is non-expansive, so $\|z_{i+1} - z_{i+2}\| = \|Tz_i - Tz_{i+1}\| \le \|z_i - z_{i+1}\|$: the residuals are non-increasing, and $(k+1)\|z_k - z_{k+1}\|^2 \le \|z_0 - z^\star\|^2$. Finally $z_{k+1} - z_k = y_{k+1} - x_{k+1}$ by the previous remark. □

3.9 Equivalence of ADMM and Dual Douglas-Rachford

ADMM is Douglas-Rachford splitting applied to the dual problem . We prove the direction used in practice: the ADMM iterates generate a DRS sequence for the dual.

The Lagrangian of $\min_{x,z} f(x)+g(z)$ subject to $Ax+Bz = c$ is $f(x)+g(z)+\langle u, Ax+Bz-c\rangle$. Its infimum over x and z is $-f^*(-A^Tu) - g^*(-B^Tu) - \langle u, c\rangle$, so the dual problem reads

$$\min_u \ \varphi(u) + \psi(u)\,, \qquad \varphi(u) = f^*(-A^Tu)\,, \qquad \psi(u) = g^*(-B^Tu) + \langle c, u\rangle\,. \tag{3.35}$$

Both functions are convex and lower semi-continuous. We use one inclusion: if $w \in \partial f^*(-A^Tu)$, then $-Aw \in \partial\varphi(u)$, since for all u', $f^*(-A^Tu') \ge f^*(-A^Tu) + \langle w, -A^T(u'-u)\rangle = \varphi(u) + \langle -Aw, u'-u\rangle$. The same holds for g^* and B. We also use that $\eta \in \partial f(x)$ implies $x \in \partial f^*(\eta)$: by the equality case of Proposition 3.4, $f(x)+f^*(\eta) = \langle x, \eta\rangle$, and for all η', $f^*(\eta') \ge \langle x, \eta'\rangle - f(x) = f^*(\eta) + \langle x, \eta'-\eta\rangle$.

Theorem 3.24 (ADMM is dual Douglas-Rachford)**.** *Let (x^k, z^k, u^k) be ADMM iterates with parameter $\rho > 0$ (the minimizers are assumed to exist), and set $\zeta_k = u^k + \rho Ax^{k+1}$. Consider DRS for $\min_u \psi(u) + \varphi(u)$ with step ρ, taking the proximal step on ψ first:*

$$v^{k+1/2} = \mathbf{prox}_{\psi,\rho}(\zeta_k)\,, \tag{3.36}$$

$$v^{k+1} = \mathbf{prox}_{\varphi,\rho}(2v^{k+1/2} - \zeta_k)\,, \tag{3.37}$$

$$\zeta_{k+1} = \zeta_k + v^{k+1} - v^{k+1/2}\,. \tag{3.38}$$

Then, for every $k \ge 0$: $v^{k+1/2} = u^{k+1}$, and the point ζ_{k+1} given by (3.38) equals $u^{k+1} + \rho Ax^{k+2}$. In other words, (ζ_k) is the DRS sequence started at $\zeta_0 = u^0 + \rho Ax^1$, and its first proximal points are the ADMM multipliers.

Proof. Recall that $p = \mathbf{prox}_{h,\rho}(\zeta)$ if and only if $(\zeta - p)/\rho \in \partial h(p)$, and that this point is unique ((3.6); uniqueness holds as soon as h is convex and $\partial h(p) \neq \emptyset$, by strict convexity of $h + \frac{1}{2\rho}\|\cdot - \zeta\|^2$).

Step 1: $v^{k+1/2} = u^{k+1}$. The z-update minimizes $g(z) + \langle u^k, Bz\rangle + \frac{\rho}{2}\|Ax^{k+1} + Bz - c\|^2$. By Fermat's rule and the sum rule, $0 \in \partial g(z^{k+1}) + B^T(u^k + \rho(Ax^{k+1} + Bz^{k+1} - c)) = \partial g(z^{k+1}) + B^Tu^{k+1}$. So $-B^Tu^{k+1} \in \partial g(z^{k+1})$, hence $z^{k+1} \in \partial g^*(-B^Tu^{k+1})$ and $-Bz^{k+1} + c \in \partial\psi(u^{k+1})$. On the other hand,

$$\frac{\zeta_k - u^{k+1}}{\rho} = \frac{u^k + \rho Ax^{k+1} - u^k - \rho(Ax^{k+1} + Bz^{k+1} - c)}{\rho} = -Bz^{k+1} + c\,.$$

Hence $u^{k+1} = \mathbf{prox}_{\psi,\rho}(\zeta_k) = v^{k+1/2}$.

Step 2: the proximal step on φ. The next x-update, $x^{k+2} \in \operatorname{argmin}_x \mathcal{L}_\rho(x, z^{k+1}, u^{k+1})$, satisfies $0 \in \partial f(x^{k+2}) + A^T\tilde{v}$ with $\tilde{v} = u^{k+1} + \rho(Ax^{k+2} + Bz^{k+1} - c)$. So $x^{k+2} \in \partial f^*(-A^T\tilde{v})$ and $-Ax^{k+2} \in \partial\varphi(\tilde{v})$. Using $u^{k+1} - u^k = \rho(Ax^{k+1} + Bz^{k+1} - c)$,

$$2u^{k+1} - \zeta_k - \tilde{v} = u^{k+1} - u^k - \rho Ax^{k+1} - \rho(Ax^{k+2} + Bz^{k+1} - c) = -\rho Ax^{k+2}\,.$$

Hence $\tilde{v} = \mathbf{prox}_{\varphi,\rho}(2v^{k+1/2} - \zeta_k) = v^{k+1}$.

Step 3: the update of ζ. By (3.38),

$$\zeta_{k+1} = u^k + \rho A x^{k+1} + \tilde{v} - u^{k+1} = u^k + \rho(Ax^{k+1} + Bz^{k+1} - c) + \rho A x^{k+2} = u^{k+1} + \rho A x^{k+2}\,.$$

This is the claimed form at index $k+1$, and the induction continues. □

Consequently, if the DRS operator for the dual has a fixed point, Proposition 3.23 applies to ζ_k: $\|\zeta_{k+1} - \zeta_k\|^2 = O(1/k)$. A linear rate for ADMM needs further assumptions (for instance strong convexity and smoothness of one of the dual functions); it is not proved in these notes.

3.10 Primal-Dual Splitting (Chambolle-Pock)

In ADMM, the update of the primal variables often requires solving a linear system (with matrix A^TA plus a multiple of the identity), which can be expensive for large problems. *Primal-dual splitting methods* only use products by the linear operator and its transpose. Consider the problem

$$\min_{x\in\mathbb{R}^n} f(x) + g(Kx)\,, \tag{3.39}$$

where f and g are proper, closed, convex functions, and $K \in \mathbb{R}^{m\times n}$. Here f has an easy proximal operator (e.g., a regularizer) and g is accessed through its conjugate g^*.

For g proper, closed and convex, $g = g^{**}$ (Fenchel-Moreau theorem, admitted) , so $g(Kx) = \sup_y \langle Kx, y\rangle - g^*(y)$, and the problem is the saddle point problem

$$\min_{x\in\mathbb{R}^n} \sup_{y\in\mathbb{R}^m} \mathcal{L}(x,y)\,, \qquad \mathcal{L}(x,y) = \langle Kx, y\rangle + f(x) - g^*(y)\,. \tag{3.40}$$

The function $\mathcal{L}$ is convex in the primal variable x and concave in the dual variable y (Figure 8). A saddle point $(x^\star, y^\star)$ satisfies $\mathcal{L}(x^\star, y) \le \mathcal{L}(x^\star, y^\star) \le \mathcal{L}(x, y^\star)$ for all (x, y).

The *Chambolle-Pock algorithm* alternates a proximal step in the primal variable and a proximal step in the dual variable, evaluated at an extrapolated primal point:

$$x_{k+1} = \mathbf{prox}_{f,\tau}(x_k - \tau K^T y_k)\,, \tag{3.41}$$

$$y_{k+1} = \mathbf{prox}_{g^*,\sigma}(y_k + \sigma K(2x_{k+1} - x_k))\,. \tag{3.42}$$

Remark 3.25 (Order of the updates and step sizes). Chambolle and Pock write the dual step first: $y_{k+1} = \mathbf{prox}_{g^*,\sigma}(y_k + \sigma K\bar{x}_k)$, $x_{k+1} = \mathbf{prox}_{f,\tau}(x_k - \tau K^T y_{k+1})$, $\bar{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$, with $\theta = 1$ and $\bar{x}_0 = x_0$. Substituting shows that $(x_k, y_{k+1})_{k\ge 0}$ then follows the updates above, started from (x_0, y_1): the two orders produce the same sequence, up to a shift of the dual index. The ergodic bound below holds under $\tau\sigma\|K\|^2 \le 1$, where $\|K\|$ is the spectral norm. Convergence of the iterates themselves is proved by Chambolle and Pock for $\tau\sigma\|K\|^2 < 1$ (admitted).

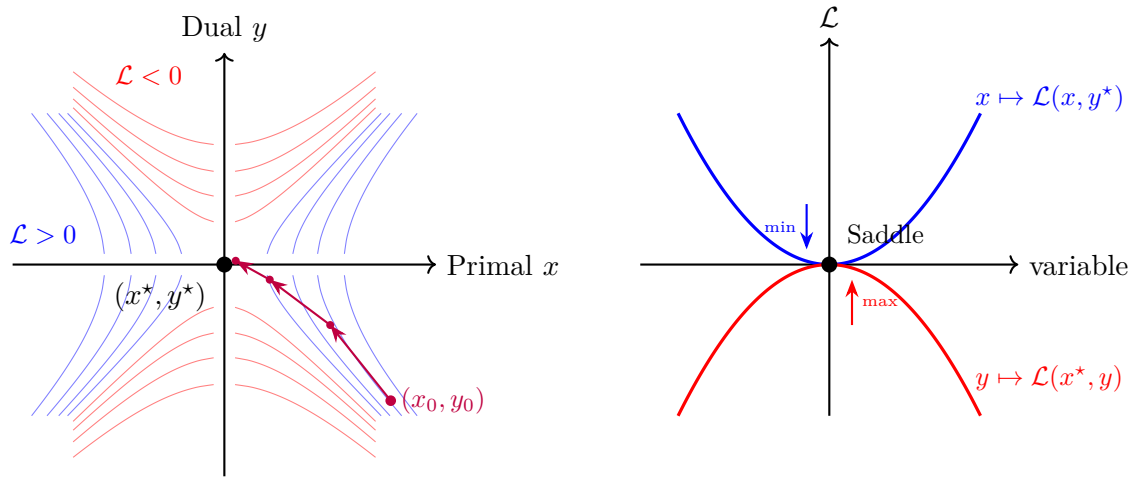
Proposition 3.26 (Moreau's Identity, admitted)**.** *Let g be a proper, closed, convex function and g^* be its conjugate. For any $\sigma > 0$ and $y \in \mathbb{R}^n$, the proximal operators of g and g^* are related by:*

$$\mathbf{prox}_{g^*,\sigma}(y) = y - \sigma\mathbf{prox}_{g,1/\sigma}\left(\frac{y}{\sigma}\right)\,. \tag{3.43}$$

This identity ensures the Chambolle-Pock algorithm is entirely computable using only the proximal operators of f and g, alongside explicit matrix-vector multiplications with K and K^T.

Theorem 3.27 (Ergodic $\mathcal{O}(1/k)$ Convergence of Chambolle-Pock)**.** *Let the step sizes satisfy $\tau\sigma\|K\|^2 \le 1$. Let (x_i, y_i) be the sequence generated by the algorithm above, initialized at (x_0, y_0), and define the ergodic averages*

$$X_k = \frac{1}{k}\sum_{i=1}^k x_i, \quad Y_k = \frac{1}{k}\sum_{i=1}^k y_i\,.$$



(a) Contour lines of $\mathcal{L}(x, y)$ *(b)* Cross-sections: convex in x, concave in y

Figure 8: Geometry of a saddle point problem, drawn for the model $\mathcal{L}(x, y) = x^2 - y^2$ (schematic). *(a)* Level curves of $\mathcal{L}$ (hyperbolas; blue: positive levels, red: negative levels) and a schematic primal-dual trajectory (purple), not computed. *(b)* One-dimensional cross-sections: fixing $y = y^\star$ gives a convex function of x (blue), fixing $x = x^\star$ gives a concave function of y (red). The saddle point minimizes the first and maximizes the second.

Then, for any $(x, y) \in \mathbf{dom}(f) \times \mathbf{dom}(g^*)$ *and* $k \geq 1$*, the restricted primal-dual gap satisfies*

$$\mathcal{L}(X_k, y) - \mathcal{L}(x, Y_k) \leq \frac{1}{k}\left(\frac{\|x - x_0\|_2^2}{\tau} + \frac{\|y - y_0\|_2^2}{\sigma}\right).$$

Proof. Write $w = (x, y)$, $w_i = (x_i, y_i)$, and

$$M = \begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}, \qquad \|w\|_M^2 = \langle w, Mw \rangle = \frac{\|x\|^2}{\tau} + \frac{\|y\|^2}{\sigma} - 2\langle Kx, y \rangle .$$

Step 1: $M \succeq 0$. By Cauchy-Schwarz and $\|K\| \leq 1/\sqrt{\tau\sigma}$, $2|\langle Kx, y\rangle| \leq \frac{2\|x\|\|y\|}{\sqrt{\tau\sigma}} \leq \frac{\|x\|^2}{\tau} + \frac{\|y\|^2}{\sigma}$. Hence $0 \leq \|w\|_M^2 \leq 2(\frac{\|x\|^2}{\tau} + \frac{\|y\|^2}{\sigma})$.

Step 2: optimality conditions. By (3.6), the two proximal steps give

$$p = \tfrac{1}{\tau}(x_i - x_{i+1}) - K^T y_i \in \partial f(x_{i+1}), \qquad q = \tfrac{1}{\sigma}(y_i - y_{i+1}) + K(2x_{i+1} - x_i) \in \partial g^*(y_{i+1}).$$

Let $(a, b) = M(w_i - w_{i+1})$, that is $a = \frac{1}{\tau}(x_i - x_{i+1}) - K^T(y_i - y_{i+1})$ and $b = \frac{1}{\sigma}(y_i - y_{i+1}) - K(x_i - x_{i+1})$. Then $p = a - K^T y_{i+1}$ and $q = b + Kx_{i+1}$.

Step 3: one-step gap. By the subgradient inequalities, $f(x) \geq f(x_{i+1}) + \langle p, x - x_{i+1}\rangle$ and $g^*(y) \geq g^*(y_{i+1}) + \langle q, y - y_{i+1}\rangle$. Hence

$$\begin{aligned}
&\mathcal{L}(x_{i+1}, y) - \mathcal{L}(x, y_{i+1}) \\
&\quad = \langle Kx_{i+1}, y\rangle - \langle Kx, y_{i+1}\rangle + f(x_{i+1}) - f(x) + g^*(y_{i+1}) - g^*(y) \\
&\quad \leq \langle Kx_{i+1}, y\rangle - \langle Kx, y_{i+1}\rangle - \langle a - K^T y_{i+1}, x - x_{i+1}\rangle - \langle b + Kx_{i+1}, y - y_{i+1}\rangle \\
&\quad = \langle a, x_{i+1} - x\rangle + \langle b, y_{i+1} - y\rangle = \langle M(w_i - w_{i+1}), w_{i+1} - w\rangle ,
\end{aligned}$$

since the six terms in K cancel: $\langle Kx_{i+1}, y\rangle - \langle Kx, y_{i+1}\rangle + \langle y_{i+1}, Kx\rangle - \langle y_{i+1}, Kx_{i+1}\rangle - \langle Kx_{i+1}, y\rangle + \langle Kx_{i+1}, y_{i+1}\rangle = 0$. For the symmetric matrix M, $2\langle M(u - v), v - w\rangle = \|u - w\|_M^2 - \|u - v\|_M^2 - \|v - w\|_M^2$. With $u = w_i$, $v = w_{i+1}$ and $M \succeq 0$:

$$\mathcal{L}(x_{i+1}, y) - \mathcal{L}(x, y_{i+1}) \leq \frac{1}{2}\left(\|w_i - w\|_M^2 - \|w_{i+1} - w\|_M^2\right).$$

Step 4: summation. Summing from $i = 0$ to $k-1$, the right-hand side telescopes and is at most $\frac{1}{2}\|w_0 - w\|_M^2$. Since $\mathcal{L}(\cdot, y)$ is convex and $\mathcal{L}(x, \cdot)$ is concave, Jensen's inequality gives

$$\mathcal{L}(X_k, y) - \mathcal{L}(x, Y_k) \leq \frac{1}{k}\sum_{i=1}^{k}\left(\mathcal{L}(x_i, y) - \mathcal{L}(x, y_i)\right) \leq \frac{1}{2k}\|w_0 - w\|_M^2 \leq \frac{1}{k}\left(\frac{\|x - x_0\|^2}{\tau} + \frac{\|y - y_0\|^2}{\sigma}\right),$$

by Step 1. □

Remark 3.28. For the dual-first order with $\tau\sigma\|K\|^2 < 1$, Chambolle and Pock obtain the bound with the smaller constants $\frac{1}{2\tau}$ and $\frac{1}{2\sigma}$ (admitted) . The proof above, a proximal point step in the metric M, gives the constants $\frac{1}{\tau}$ and $\frac{1}{\sigma}$.

3.11 Accelerated Chambolle-Pock for Strongly Convex Objectives

When f is γ-strongly convex ($\gamma > 0$; for a non-differentiable f, this means that $f - \frac{\gamma}{2}\|\cdot\|^2$ is convex), the Chambolle-Pock algorithm can be accelerated to reach the rate $O(1/k^2)$ on the primal distance $\|x_k - x^\star\|_2^2$. The step sizes τ_k, σ_k and the relaxation parameter θ_k change at each iteration. The algorithm below is written in the dual-first order of Chambolle and Pock.

Algorithm 3 Accelerated Primal-Dual Algorithm (Strongly Convex f)

1: *Initialization:* Choose $\tau_0, \sigma_0 > 0$ such that $\tau_0\sigma_0\|K\|^2 \leq 1$. Set $(x_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^m$ and $\bar{x}_0 = x_0$.
2: **for** $k \geq 0$ **do**
3: *Dual Step:* $y_{k+1} = \mathrm{prox}_{g^*,\sigma_k}(y_k + \sigma_k K\bar{x}_k)$
4: *Primal Step:* $x_{k+1} = \mathrm{prox}_{f,\tau_k}(x_k - \tau_k K^T y_{k+1})$
5: *Parameter Updates:*

$$\theta_k = \frac{1}{\sqrt{1 + 2\gamma\tau_k}}, \quad \tau_{k+1} = \theta_k\tau_k, \quad \sigma_{k+1} = \sigma_k/\theta_k$$

6: *Extrapolation:* $\bar{x}_{k+1} = x_{k+1} + \theta_k(x_{k+1} - x_k)$
7: **end for**

Theorem 3.29 ($O(1/k^2)$ Accelerated Convergence)**.** *Let f be γ-strongly convex, let $(x^\star, y^\star)$ be a saddle point of $\mathcal{L}$, and run the accelerated algorithm with $\tau_0\sigma_0\|K\|^2 \leq 1$.*

1. *For all k, $\tau_k\sigma_k = \tau_0\sigma_0$, $\frac{1}{\tau_{k+1}^2} = \frac{1+2\gamma\tau_k}{\tau_k^2}$, and for $k \geq 1$,*

$$\tau_k \leq \frac{2 + \gamma\tau_0}{2\gamma k}.$$

2. *(Admitted.) There is a constant C_0, depending only on τ_0, σ_0, $\|x_0 - x^\star\|$ and $\|y_0 - y^\star\|$, such that $\|x_k - x^\star\|^2 \leq C_0\,\tau_k^2$ for all k .*

3. *Consequently, for $k \geq 1$, $\|x_k - x^\star\|_2^2 \leq \dfrac{C}{\gamma^2 k^2}$ with $C = \frac{C_0(2+\gamma\tau_0)^2}{4}$.*

Proof. Item 1. $\tau_{k+1}\sigma_{k+1} = (\theta_k\tau_k)(\sigma_k/\theta_k) = \tau_k\sigma_k$. From $\tau_{k+1} = \tau_k/\sqrt{1 + 2\gamma\tau_k}$, $\tau_{k+1}^2 = \tau_k^2/(1 + 2\gamma\tau_k)$, which is the second identity. Let $a_k = 1/\tau_k$, so that $a_{k+1}^2 = a_k^2 + 2\gamma a_k$. Then $a_{k+1} \leq a_k + \gamma$ and a_k is increasing, so

$$a_{k+1} - a_k = \frac{2\gamma a_k}{a_{k+1} + a_k} \geq \frac{2\gamma a_k}{2a_k + \gamma} \geq \frac{2\gamma a_0}{2a_0 + \gamma},$$

where the last step uses that $s \mapsto 2\gamma s/(2s+\gamma)$ is increasing. Hence $a_k \geq a_0 + k\frac{2\gamma a_0}{2a_0+\gamma} \geq k\frac{2\gamma a_0}{2a_0+\gamma}$, that is $\tau_k \leq \frac{2a_0+\gamma}{2\gamma a_0 k} = \frac{2+\gamma\tau_0}{2\gamma k}$.

Item 3. Combine items 1 and 2: $\|x_k - x^\star\|^2 \leq C_0\tau_k^2 \leq C_0\frac{(2+\gamma\tau_0)^2}{4\gamma^2k^2}$. □

Remark 3.30. Item 1 also shows $\gamma k\tau_k \to 1$: $a_{k+1} - a_k \to \gamma$ since $a_k \to \infty$. So τ_k decreases like $1/(\gamma k)$, not like $1/k^2$, and the rate on $\|x_k - x^\star\|^2$ comes from the factor τ_k^2 in item 2. In the proof of Chambolle and Pock, the energy is weighted by $1/\tau_k^2$ for the primal distance and by $1/(\tau_0\sigma_0)$ for the dual distance, and the cross terms created by the extrapolation must be handled; this part is admitted here. The primal-dual gap is not covered by this theorem.

4 Second-Order Methods

First-order methods rely on linear approximations. Second-order methods use curvature information (the Hessian $\nabla^2 f(x)$) and the local quadratic model

$$q_x(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x)\,. \tag{4.1}$$

If $\nabla^2 f(x) \succ 0$, minimizing q_x gives the Newton step.

4.1 Newton's Method

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k) \tag{4.2}$$

Unlike gradient descent, Newton's method is *affine invariant*: for an invertible matrix A, Newton's method applied to $h(y) = f(Ay)$ from $y_0 = A^{-1}x_0$ gives $y_k = A^{-1}x_k$ (since $\nabla h(y) = A^T\nabla f(Ay)$ and $\nabla^2 h(y) = A^T\nabla^2 f(Ay)A$). Its iterates do not depend on a linear change of coordinates, which is a source of ill-conditioning for GD.

Theorem 4.1 (Quadratic Convergence of Newton's Method)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and m-strongly convex, with minimizer $x^\star$, and assume its Hessian is M-Lipschitz continuous (i.e., $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \le M\|x-y\|$). Then the Newton iterates satisfy, for all k,*

$$\|x_{k+1} - x^\star\| \le \frac{M}{2m}\|x_k - x^\star\|^2\,. \tag{4.3}$$

If, in addition, $q = \frac{M}{2m}\|x_0 - x^\star\| < 1$, then $\|x_k - x^\star\| \le \frac{2m}{M}\,q^{2^k}$ for all $k \ge 0$ (quadratic convergence).

Proof. Let $e_k = x_k - x^\star$. Since $\nabla f(x^\star) = 0$, we subtract it from the gradient expansion. By the Fundamental Theorem of Calculus:

$$\nabla f(x_k) - \nabla f(x^\star) = \int_0^1 \nabla^2 f(x^\star + t(x_k - x^\star))(x_k - x^\star)dt$$

Recall the update rule: $x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$. Subtracting $x^\star$ from both sides:

$$e_{k+1} = e_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$$

Multiply by the Hessian $\nabla^2 f(x_k)$:

$$\nabla^2 f(x_k)e_{k+1} = \nabla^2 f(x_k)e_k - \nabla f(x_k)$$

Substituting the integral expression for $\nabla f(x_k)$ (noting $\nabla f(x^\star) = 0$):

$$\begin{aligned}\nabla^2 f(x_k)e_{k+1} &= \nabla^2 f(x_k)e_k - \int_0^1 \nabla^2 f(x^\star + te_k)e_k dt \\ &= \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^\star + te_k)\right)e_k dt\end{aligned}$$

Taking norms and using the M-Lipschitz property of the Hessian:

$$\begin{aligned}\|\nabla^2 f(x_k)e_{k+1}\| &\le \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^\star + te_k)\|\|e_k\|dt \\ &\le \int_0^1 M\|x_k - (x^\star + te_k)\|\|e_k\|dt \\ &= \int_0^1 M(1-t)\|e_k\|^2 dt = \frac{M}{2}\|e_k\|^2\end{aligned}$$

Finally, since f is m-strongly convex, $\nabla^2 f(x_k) \succeq mI$. So the Hessian $\nabla^2 f(x_k)$ is invertible, and its inverse has operator norm at most $1/m$. Thus:

$$\|e_{k+1}\| \leq \|\nabla^2 f(x_k)^{-1}\|_{op} \|\nabla^2 f(x_k) e_{k+1}\| \leq \frac{1}{m} \cdot \frac{M}{2} \|e_k\|^2 .$$

For the second claim, let $r_k = \frac{M}{2m}\|e_k\|$. The first claim reads $r_{k+1} \leq r_k^2$, so $r_k \leq r_0^{2^k} = q^{2^k}$ by induction. □

Remark 4.2 (Session 3 notebook, experiment 4). On ℓ_2-regularized logistic regression ($N = 200$, $d = 10$, $\kappa = 2.84$), the gaps $f(x_k) - f^\star$ of Newton's method are $1.8 \cdot 10^{-1}$, $1.1 \cdot 10^{-3}$, $2.6 \cdot 10^{-7}$, $1.4 \cdot 10^{-14}$ for $k = 0, \ldots, 3$, and below 10^{-15} from $k = 4$. GD with step $1/L$ reaches $9.0 \cdot 10^{-6}$ after 5 iterations and $1.3 \cdot 10^{-9}$ after 10 (`figures/fig_td3_newton_gd.pdf`).

4.2 Quasi-Newton Methods (BFGS)

In high dimension, computing and inverting the Hessian ($\mathcal{O}(n^3)$ operations) is too expensive. *Quasi-Newton methods* use approximations $B_k \approx \nabla^2 f(x_k)$ built from gradient differences only, which satisfy the *secant equation*:

$$B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k) \tag{4.4}$$

The most used method, *BFGS* (Broyden-Fletcher-Goldfarb-Shanno), updates the inverse Hessian approximation directly with rank-2 updates costing $\mathcal{O}(n^2)$. *L-BFGS* (limited-memory BFGS) stores only the last p pairs of iterate and gradient differences, for a cost $\mathcal{O}(pn)$ per iteration. It is, for instance, the default solver of logistic regression in scikit-learn.

Let $s_k = x_{k+1} - x_k$. The quasi-Newton step with unit step length is $x_{k+1} = x_k - B_k^{-1}\nabla f(x_k)$. The following two results are admitted .

Theorem 4.3 (Dennis-Moré, admitted)**.** *Let f be twice continuously differentiable with $\nabla^2 f(x^\star) \succ 0$ at a point $x^\star$ with $\nabla f(x^\star) = 0$. Let $x_{k+1} = x_k - B_k^{-1}\nabla f(x_k)$, with B_k invertible, $x_k \neq x^\star$ and $x_k \to x^\star$. Then*

$$\|x_{k+1} - x^\star\| = o(\|x_k - x^\star\|) \quad \textit{if and only if} \quad \frac{\|(B_k - \nabla^2 f(x^\star))\, s_k\|}{\|s_k\|} \to 0 .$$

Theorem 4.4 (Convergence of BFGS and L-BFGS, admitted)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and strongly convex, with $\nabla^2 f$ Lipschitz continuous near the minimizer $x^\star$. BFGS with a line search satisfying the Wolfe conditions, in which the unit step is tried first, converges Q-superlinearly to $x^\star$. L-BFGS with memory $p \geq 1$ converges R-linearly: there are $C > 0$ and $r \in (0, 1)$ with $f(x_k) - f^\star \leq Cr^k$.*

Remark 4.5 (Superlinear Convergence Without Hessian Convergence). The Dennis-Moré condition does *not* require the matrices B_k to converge to $\nabla^2 f(x^\star)$: it suffices that B_k approximates $\nabla^2 f(x^\star)$ *along the steps* s_k. Full BFGS achieves this. For L-BFGS with a fixed memory, only R-linear convergence is guaranteed by Theorem 4.4; in practice it is often much faster than gradient descent (next remark).

Remark 4.6 (Session 3 notebook, experiment 5). On a quadratic of dimension 1000 with eigenvalues $1, \ldots, 100$ ($\kappa = 100$), L-BFGS (SciPy's L-BFGS-B, memory 10) reaches a gap below 10^{-8} at iteration 59, while GD with step $1/L$ is at 0.23 after 100 iterations and $1.3 \cdot 10^{-5}$ after 500 (`figures/fig_td3_lbfgs.pdf`).

4.3 Interior Point Methods (IPM)

Consider the convex problem $\min f_0(x)$ subject to $f_i(x) \leq 0$, $i = 1, \dots, m$, with f_0, f_i convex and twice differentiable. Barrier methods solve a sequence of unconstrained problems, using the *logarithmic barrier*:

$$\min_x \phi_t(x) = t f_0(x) - \sum_{i=1}^{m} \log(-f_i(x)), \qquad t > 0, \tag{4.5}$$

defined on the strictly feasible set $\{x : f_i(x) < 0 \ \forall i\}$. The term $-\log(-f_i(x))$ tends to $+\infty$ at the boundary and keeps the iterates in the interior.

Definition 4.7 (The Central Path)**.** Assume that ϕ_t has a unique minimizer $x^\star(t)$ for every $t > 0$. The central path is the curve $t \mapsto x^\star(t)$.

Proposition 4.8 (Duality gap on the central path)**.** *Under the assumptions above,* $f_0(x^\star(t)) - p^\star \leq m/t$.

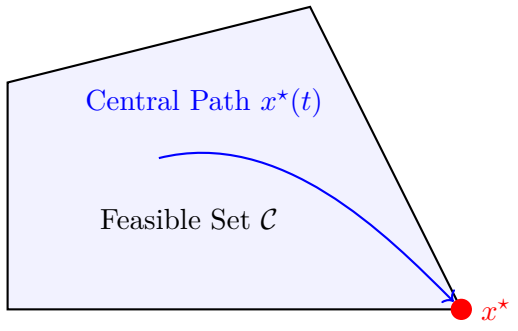
Proof. At $x = x^\star(t)$, $\nabla \phi_t(x) = t\nabla f_0(x) + \sum_i \frac{\nabla f_i(x)}{-f_i(x)} = 0$. Let $\lambda_i = \frac{1}{-t f_i(x)} > 0$. Dividing by t, $\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) = 0$, so x minimizes the convex function $\mathcal{L}(\cdot, \lambda)$ and $g(\lambda) = f_0(x) + \sum_i \lambda_i f_i(x) = f_0(x) - \frac{m}{t}$. By weak duality, $p^\star \geq g(\lambda) = f_0(x^\star(t)) - \frac{m}{t}$. □

In particular $f_0(x^\star(t)) \to p^\star$ as $t \to \infty$.

The barrier method:

1. Initialize $t = t_0 > 0$, a strictly feasible x, and $\mu > 1$.
2. *Centering step:* compute $x^\star(t)$ by minimizing ϕ_t with Newton's method, started from the current x.
3. Stop if $m/t < \epsilon$; then $f_0(x^\star(t)) - p^\star < \epsilon$ by Proposition 4.8.
4. *Update:* $t \leftarrow \mu t$, and go to step 2.

At the j-th centering step, $t = \mu^{j-1} t_0$, so the number of centering steps is at most $\max\left(1, 2 + \log(m/(\epsilon t_0))/\log \mu\right)$, that is $O(\log(1/\epsilon))$. The number of Newton steps per centering step is bounded for self-concordant barriers (admitted) .



Path following:
For small t, $x^\star(t)$ is close to the analytic center (far from the boundary).
As $t \to \infty$, $f_0(x^\star(t)) \to p^\star$.

Figure 9: Interior point methods (schematic): instead of moving along the boundary like the simplex method, IPMs follow the central path through the interior.

5 Numerical Experiments

The two benchmarks below are computed in the session 3 notebook (`td3.ipynb`, experiments 6 and 7), with one process, one thread and fixed random seeds. The session notebooks also contain the smaller experiments cited in the remarks of the previous sections.

Data. Both benchmarks use synthetic data of the size of the California Housing dataset (which is not used): $n = 20\,640$ observations, $d = 8$ features. The features are Gaussian, with correlation $0.9^{|i-j|}$ between features i and j, then centered and scaled; $y = Xw^\natural + \varepsilon$ with $w^\natural = (2, -1, 0, 0, 1.5, 0, 0, 0.5)$ and $\varepsilon \sim \mathcal{N}(0, I_n)$. The conclusions below hold for these data only. Both figures count epochs or iterations, not wall-clock time.

5.1 Benchmark 1: Ridge Regression

We minimize $f(w) = \frac{1}{2n}\|y - Xw\|^2 + \frac{\mu}{2}\|w\|^2$ with $\mu = 10^{-3}$. Here $L = 6.2041$, $m = 0.0562$, $\kappa = 110.4$, the largest smoothness constant of the individual terms is $L_{\max} = 118.64$, and $f(0) - f^\star = 3.5558$. The horizontal axis counts epochs (number of evaluations of gradients ∇f_i, divided by n), with a budget of 50 epochs. Step sizes: GD $1/L$; Nesterov's method with step $1/L$ and constant momentum $(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ (the variant for strongly convex functions, not analyzed in these notes); SGD $\min(1/(2L_{\max}), 1/(m(k+1)))$; SVRG $\eta = 1/(4L_{\max})$ with $M = n$; L-BFGS (SciPy's L-BFGS-B, memory 10); Newton's method takes one step, counted as $d + 1 = 9$ epochs and plotted as an isolated point.

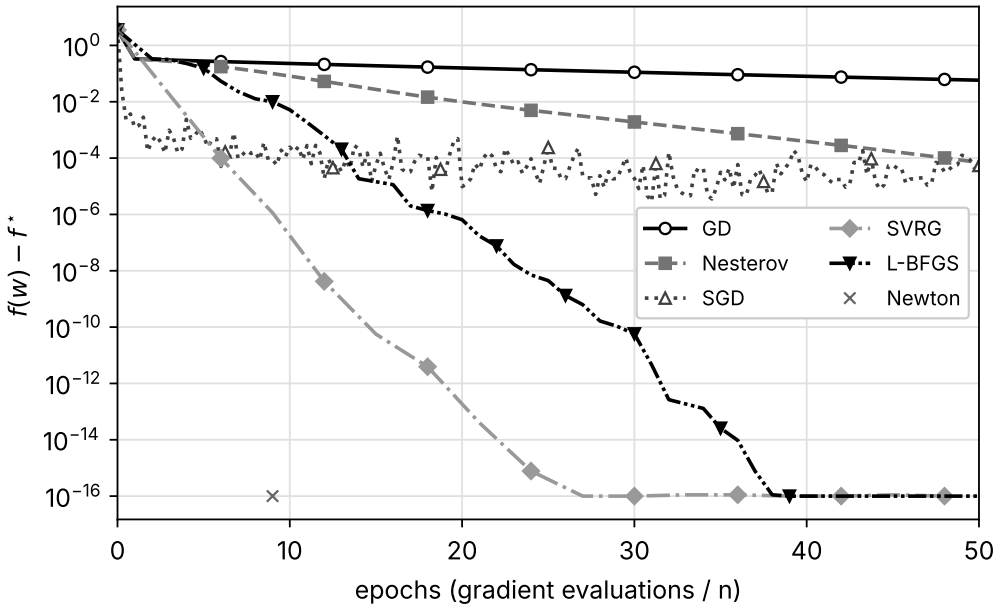


Figure 10: Ridge regression on synthetic data ($n = 20\,640$, $d = 8$, $\kappa = 110.4$): gap $f(w) - f^\star$ against epochs. Gaps are clipped at 10^{-16} for the log scale.

Remark 5.1 (Reading Figure 10). 1. The objective is quadratic, so one Newton step reaches the minimizer (the quadratic model q_x is f itself).

2. With $\kappa = 110$, GD decreases slowly: its gap is $5.8 \cdot 10^{-2}$ after 50 epochs, and Nesterov's method is at $7.0 \cdot 10^{-5}$.

3. SGD decreases quickly in the first epochs. After that, its gap fluctuates between about 10^{-5} and 10^{-3}, and it is $5.5 \cdot 10^{-5}$ after 50 epochs; at this scale, the figure does not show whether it still decreases.

Method	Gap at the last point (epochs)	First epoch count with gap $< 10^{-10}$
GD	$5.8 \cdot 10^{-2}$ (50)	not reached
Nesterov	$7.0 \cdot 10^{-5}$ (50)	not reached
SGD	$5.5 \cdot 10^{-5}$ (50)	not reached
SVRG	$< 10^{-15}$ (48)	15
L-BFGS	$< 10^{-15}$ (50)	30
Newton	$< 10^{-15}$ (9)	9

Table 1: Values read from the notebook output for Figure 10.

4. SVRG and L-BFGS reach a gap below 10^{-10} after 15 and 30 epochs. The SVRG step $1/(4L_{\max})$ is outside the range of Theorem 2.12 ($\rho \geq 1$ for $\eta = 1/(4L_{\max})$): the linear convergence is observed, not guaranteed by the theorem.

5.2 Benchmark 2: LASSO

We minimize $F(w) = \frac{1}{2n}\|y - Xw\|^2 + \lambda\|w\|_1$ on the same data, with $\lambda = \lambda_{\max}/10 = 0.2434$, where $\lambda_{\max} = \|X^T y\|_\infty / n = 2.4337$ is the smallest λ for which $w = 0$ is a solution (Proposition 1.26, after scaling by $1/n$). Here $L = 6.2031$ and $F^\star = 1.2707334507$. The solution is $w^\star = (1.1188, 0, 0, 0, 1.196, 0, 0, 0.3621)$: its support $\{1, 5, 8\}$ is the support of $w^\natural$ without its second coordinate. We run 100 iterations of ISTA and FISTA (step $1/L$), cyclic coordinate descent (one iteration is one sweep over the d coordinates) and ADMM ($\rho = 1$). The horizontal axis counts iterations, not time. ISTA, FISTA and ADMM use the Gram matrix $X^T X$, computed once, so that one of their iterations costs $O(d^2)$.

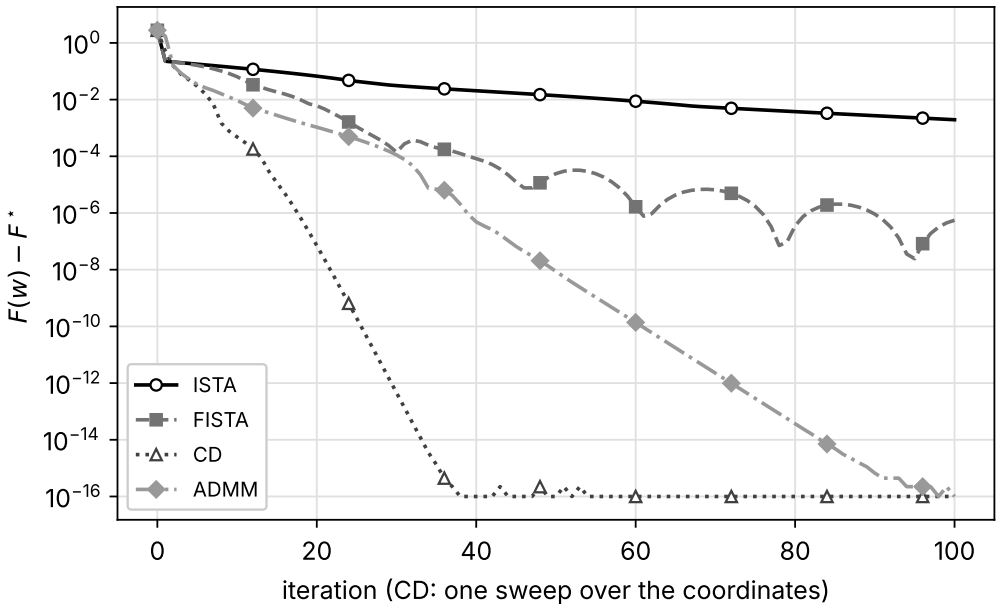


Figure 11: LASSO on the same synthetic data, $\lambda = \lambda_{\max}/10$: gap $F(w) - F^\star$ against iterations (for CD, one sweep over the coordinates). Gaps are clipped at 10^{-16} for the log scale.

Remark 5.2 (Reading Figure 11). 1. FISTA is faster than ISTA, which is consistent with the bounds of Theorems 3.7 and 3.10, but its gap is not monotone: it oscillates.

2. On this small problem ($d = 8$), CD reaches a gap below 10^{-10} after 26 sweeps and ADMM after 61 iterations. No rate for CD or ADMM is proved in these notes.

Method	Gap after 10 / 50 / 100 iterations	First iteration with gap $< 10^{-10}$
ISTA	$1.3 \cdot 10^{-1}$ / $1.4 \cdot 10^{-2}$ / $1.9 \cdot 10^{-3}$	not reached
FISTA	$6.4 \cdot 10^{-2}$ / $2.4 \cdot 10^{-5}$ / $5.6 \cdot 10^{-7}$	not reached
CD	$4.8 \cdot 10^{-4}$ / $< 10^{-15}$ / $< 10^{-15}$	26
ADMM	$9.3 \cdot 10^{-3}$ / $8.7 \cdot 10^{-9}$ / $< 10^{-15}$	61

Table 2: Values read from the notebook output for Figure 11.

3. The comparison is per iteration; the cost of one iteration differs between methods. These rankings are for one data set and one λ; they are not general rules.

6 Summary of Main Results

We collect the main convergence results of these notes. "Admitted" marks results stated without proof.

Convergence Rate Comparison

Table 3: Convergence results of the main algorithms. Here $\kappa = L/m$, SC = strongly convex, x_0 is the starting point, $x^\star$ a minimizer, $R \geq \|x_0 - x^\star\|$. The cost column counts oracle calls; for a finite sum of n terms in dimension d, one full gradient costs $O(nd)$ and one ∇f_i costs $O(d)$.

Algorithm	Assumptions	Guarantee	Cost per iteration
Deterministic smooth methods			
GD, $t = \frac{1}{L}$	convex, L-smooth	$f(x_k) - f^\star \leq \frac{L\|x_0 - x^\star\|^2}{2k}$ (Theorem 2.1)	1 gradient
GD, $t = \frac{1}{L}$	m-SC, L-smooth	$f(x_k) - f^\star$ and $\|x_k - x^\star\|^2$ decrease by $(1 - \frac{1}{\kappa})$ per step (Theorem 2.1)	1 gradient
Nesterov (FISTA with $g = 0$), $t = \frac{1}{L}$	convex, L-smooth	$f(x_k) - f^\star \leq \frac{2L\|x_0 - x^\star\|^2}{(k+1)^2}$ (Theorem 3.10)	1 gradient
GD + PL, $t = \frac{1}{L}$	L-smooth, μ-PL, $\inf f > -\infty$	$f(x_k) - f^\star \leq (1 - \frac{\mu}{L})^k (f(x_0) - f^\star)$	1 gradient
Non-smooth and proximal methods			
Subgradient, horizon k	convex, G-Lipschitz	$\min_{i \leq k} f(x_i) - f^\star \leq \frac{RG}{\sqrt{k+1}}$ (Theorem 2.14)	1 subgradient
ISTA, $t \leq \frac{1}{L}$	f convex L-smooth, g convex	$F(x_k) - F^\star \leq \frac{\|x_0 - x^\star\|^2}{2tk}$ (Theorem 3.7)	1 gradient + 1 prox
FISTA, $t = \frac{1}{L}$	f convex L-smooth, g convex	$F(x_k) - F^\star \leq \frac{2L\|x_0 - x^\star\|^2}{(k+1)^2}$ (Theorem 3.10)	1 gradient + 1 prox
Operator splitting methods			
Douglas-Rachford	f, g convex, fixed point $z^\star$ exists	$\|z_{k+1} - z_k\|^2 \leq \frac{\|z_0 - z^\star\|^2}{k+1}$ (Proposition 3.23)	2 prox
Peaceman-Rachford ($\alpha = 1$), $t = \frac{1}{\sqrt{mL}}$	f m-SC and L-smooth	$\|v_k - v^\star\| \leq (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^k \|v_0 - v^\star\|$ (Theorem 3.21)	2 prox
Douglas-Rachford ($\alpha = \frac{1}{2}$), $t = \frac{1}{\sqrt{mL}}$	f m-SC and L-smooth	$\|v_k - v^\star\| \leq (\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1})^k \|v_0 - v^\star\|$ (Theorem 3.21)	2 prox

Algorithm	Assumptions	Guarantee	Cost per iteration
ADMM	linear constraints, subproblems solvable	DRS on the dual (Theorem 3.24); if the dual DRS has a fixed point, $\|\zeta_{k+1}-\zeta_k\|^2 = O(1/k)$	x- and z-minimizations
Chambolle-Pock	f, g convex, $\tau\sigma\|K\|^2 \le 1$	ergodic gap $\le \frac{1}{k}(\frac{\|x-x_0\|^2}{\tau} + \frac{\|y-y_0\|^2}{\sigma})$ (Theorem 3.27)	$\mathbf{prox}_f + \mathbf{prox}_{g^*}$, K, K^T
CP accelerated	f γ-SC	$\|x_k - x^\star\|^2 \le \frac{C}{\gamma^2 k^2}$ (Theorem 3.29, partly admitted)	$\mathbf{prox}_f + \mathbf{prox}_{g^*}$, K, K^T
Stochastic methods			
SGD, $t_k = \frac{1}{m(k+1)}$	m-SC, L-smooth, $\mathbb{E}\|\nabla f_{i_k}(x_k)\|^2 \le G^2$	$\mathbb{E}[f(x_k) - f^\star] \le \frac{LG^2}{2m^2 k}$ (Theorem 2.6)	1 ∇f_i
SGD, $t = \frac{R}{G\sqrt{T}}$, averaged	convex, $\mathbb{E}\|\nabla f_{i_k}(x_k)\|^2 \le G^2$ for every constant step	$\mathbb{E}[f(\bar{x}_T) - f^\star] \le \frac{RG}{\sqrt{T}}$ (Theorem 2.9)	1 ∇f_i
SVRG (admitted)	f_i convex L-smooth, f m-SC, $\eta < \frac{1}{4L}$, $\rho < 1$	$\mathbb{E}[f(\tilde{x}_s) - f^\star] \le \rho^s (f(\tilde{x}_0) - f^\star)$ (Theorem 2.12)	$n + 2M$ ∇f_i per epoch
Adam	none	no rate in these notes	1 ∇f_i
Second-order and constrained methods			
Newton	m-SC, C^2, M-Lipschitz Hessian	$\|x_{k+1} - x^\star\| \le \frac{M}{2m}\|x_k - x^\star\|^2$; quadratic if $\frac{M}{2m}\|x_0 - x^\star\| < 1$	gradient, Hessian, $O(d^3)$ solve
BFGS (admitted)	SC, C^2, Lipschitz Hessian near $x^\star$, Wolfe	Q-superlinear (Theorem 4.4)	$O(d^2)$ + gradient
L-BFGS (admitted)	same, memory p	R-linear (Theorem 4.4)	$O(pd)$ + gradient
Barrier method	convex, central path exists	$O(\log(1/\epsilon))$ centering steps for a gap $< \epsilon$ (Proposition 4.8)	Newton steps per centering
Structural and projection-free methods			
Coordinate descent (admitted)	f convex smooth, g separable	convergence of limit points; no rate here	1 coordinate; LASSO sweep $O(n_{\text{obs}} d)$
Frank-Wolfe, $\gamma_k = \frac{2}{k+2}$	convex, L-smooth, $\mathcal{C}$ compact, diameter D	$f(x_k) - f^\star \le \frac{2LD^2}{k+2}$ (Theorem 2.20)	1 gradient + 1 LMO

Key Messages

Table 3 shows a trade-off between *convergence speed* and *cost per iteration*. Newton's method converges quadratically near the solution, but each step requires a Hessian and a linear solve. First-order methods (GD, Nesterov, ISTA/FISTA) are cheap per iteration but converge sublinearly unless the objective is strongly convex (or PL). Stochastic methods (SGD, Adam, SVRG) use one ∇f_i per step; SGD pays for this with slower, noisier convergence, while SVRG recovers a linear rate by computing a full gradient once per epoch. Five points to remember:

1. *Strong convexity gives linear rates.* For an m-strongly convex objective, the guarantee of

Table 4: Practical guidance: which algorithm for which problem (see the guarantees in Table 3).

Scenario	Algorithms to consider
Large-scale, smooth, unconstrained	SGD or Adam (cheap steps); SVRG or L-BFGS for high accuracy
Sparse regularization (ℓ_1)	FISTA, coordinate descent (+ safe screening)
Constrained with expensive projection	Frank-Wolfe (projection-free, via the LMO)
Low-rank / nuclear norm	Frank-Wolfe (rank-one LMO), or projected gradient at small scale
Small-to-medium, smooth	Newton or L-BFGS
Non-smooth, convex, no easy prox	Subgradient method
Deep learning training	Adam/AdamW (+ learning rate schedule)
Separable penalty, large d	Coordinate descent with screening
Composite with linear coupling	ADMM, Chambolle-Pock
Sum of two non-smooth functions	Douglas-Rachford splitting

GD improves from $O(1/k)$ to the linear rate $(1 - 1/\kappa)^k$. The Polyak-Łojasiewicz condition gives the same rate without convexity.

2. *Momentum accelerates first-order methods.* Nesterov's momentum improves the guarantee from $O(1/k)$ to $O(1/k^2)$ for smooth convex problems, and from ISTA to FISTA in the composite setting, for one extra vector operation per step. FISTA is not a descent method.

3. *Structure should be used.* Sparsity (ℓ_1), separability and the geometry of the constraints guide the choice of the algorithm. Safe screening removes variables with a zero coefficient at the optimum, using only a duality gap. Frank-Wolfe replaces expensive projections (e.g., a full SVD for the nuclear norm ball) with a linear minimization oracle (a leading singular pair).

4. *No single algorithm is best.* The best solver depends on the size of the problem, the structure of the regularization and the smoothness of the objective (Table 4); the two benchmarks of Section 5 give different rankings.

5. *Splitting methods use proximal steps.* Douglas-Rachford, ADMM and Chambolle-Pock split a composite problem into proximal steps. When f is strongly convex and smooth, Douglas-Rachford and Peaceman-Rachford converge linearly, with factors governed by κ (Theorem 3.21). No linear rate is proved here for ADMM or Chambolle-Pock.

List of Notation

Symbol	*Description*
$\mathbb{R}^n$	n-dimensional Euclidean space
$\|x\|$, $\|x\|_2$	Euclidean (ℓ_2) norm
$\|x\|_1$	ℓ_1 norm: $\sum_i \|x_i\|$
$\|X\|_*$	Nuclear norm: $\sum_i \sigma_i(X)$
$\|X\|_F$	Frobenius norm: $\sqrt{\sum_{ij} X_{ij}^2}$
$\langle x, y\rangle$	Inner product $x^T y$
$\nabla f(x)$	Gradient of f at x
$\partial f(x)$	Subdifferential (set of subgradients)
$\nabla^2 f(x)$	Hessian matrix of f at x
$\mathbf{epi}(f)$	Epigraph: $\{(x,t) : f(x) \le t\}$
$\mathbf{dom}(f)$	Domain of f
$f^\star$	Optimal value: $\inf f$
$x^\star$	A minimizer of f
g^*	Fenchel conjugate of g (the star *; the PGD course writes $\star$ for the conjugate and the adjoint)
L	Smoothness constant (Lipschitz constant of ∇f)
m (or μ)	Strong convexity constant (or PL constant); in Section 1.7, m is the number of inequality constraints
n	Dimension of x; for finite sums $\frac{1}{n}\sum_i f_i$ and data matrices, n is the number of observations and d the dimension
$\kappa = L/m$	Condition number
G	Lipschitz constant of f (subgradient method) or bound on $\mathbb{E}\|\nabla f_{i_k}(x_k)\|^2 \le G^2$ (SGD)
σ^2	Variance bound for stochastic gradients
t, t_k	Step size (learning rate)
γ_k	Frank-Wolfe step size
β_1, β_2	Adam momentum decay parameters
λ	Regularization parameter (Sec. 3) or Lagrange multiplier (Sec. 1.7)
ν	Lagrange multiplier for equality constraints
ρ	ADMM penalty parameter or SVRG contraction factor
α	Relaxation parameter in generalized Douglas-Rachford
τ, σ	Primal and dual step sizes in Chambolle-Pock
θ, θ_k	Extrapolation parameter in Chambolle-Pock
s_k, β_k	Momentum sequence and coefficient in FISTA
$F(x) = f(x) + g(x)$	Composite objective: f smooth, g non-smooth (Sec. 3)
$\mathcal{L}$	Lagrangian (Section 1.7) or saddle function of Chambolle-Pock
$\iota_{\mathcal{C}}$	Indicator of $\mathcal{C}$: 0 on $\mathcal{C}$, $+\infty$ outside
$\mathbf{prox}_{g,t}(v)$	Proximal operator of g with parameter t at point v
$R_{f,t} = 2\mathbf{prox}_{f,t} - \mathrm{Id}$	Reflected proximal operator (reflected resolvent) of f
$J_{tA} = (\mathrm{Id} + tA)^{-1}$	Resolvent of a maximal monotone operator A
$g^*(y)$	Fenchel-Legendre conjugate of g
K	Linear operator in Chambolle-Pock formulation
$\mathcal{L}_\rho(x, z, u)$	Augmented Lagrangian with penalty ρ

Symbol	*Description*
$\Pi_{\mathcal{C}}(x)$	Projection of x onto convex set $\mathcal{C}$
$\mathbf{LMO}_{\mathcal{C}}(v)$	Linear Minimization Oracle: a point of $\operatorname{argmin}_{s\in\mathcal{C}}\langle s, v\rangle$
$\|\cdot\|_*$	Dual norm (also the nuclear norm $\|X\|_*$ for matrices)
$A \succeq 0$	A is positive semidefinite
$\sigma_i(X)$	i-th singular value of X
$\operatorname{rank}(X)$	Rank of matrix X
Ω	Observation set (matrix completion)
P_Ω	Projection onto observed entries

Glossary

Affine Invariance Property of an algorithm whose convergence behavior is unchanged under invertible affine transformations of the variable space. Newton's method is affine invariant; Gradient Descent is not.

Central Path The smooth trajectory $x^\star(t)$ of minimizers of the logarithmic barrier function parameterized by $t > 0$, followed by Interior Point Methods to reach the constrained optimum (Sec. 4).

Complementary Slackness KKT condition requiring $\lambda_i^\star f_i(x^\star) = 0$: at optimality, for each i, the constraint is active ($f_i(x^\star) = 0$) or its multiplier is zero ($\lambda_i^\star = 0$).

Chambolle-Pock (Primal-Dual Hybrid Gradient) Algorithm for problems of the form $\min_x f(x)+g(Kx)$ with a linear operator K. Alternates proximal steps on the primal ($\mathbf{prox}_f$) and dual ($\mathbf{prox}_{g^*}$) variables (Sec. 3).

Co-coercivity Property of an operator A: $\langle Ax - Ay, x - y\rangle \geq \beta\|Ax - Ay\|^2$ for some $\beta > 0$. The gradient of a convex L-smooth function is $(1/L)$-co-coercive (Baillon-Haddad theorem, admitted).

Composite Optimization Minimization of $F(x) = f(x)+g(x)$ where f is smooth and g is non-smooth but has a tractable proximal operator (Sec. 3).

Condition Number Ratio $\kappa = L/m$ of the smoothness constant to the strong convexity constant. The guarantees of first-order methods degrade when κ grows.

Convergence Rate The speed at which an error measure of an algorithm decreases. Common rates: *sublinear* $O(1/k)$, $O(1/k^2)$ (accelerated), *linear* ρ^k ($\rho < 1$), *superlinear*, *quadratic* (Newton).

Convex Function A function f such that $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ for all $\theta \in [0, 1]$. Equivalently: a function whose epigraph is a convex set.

Descent Lemma Inequality for L-smooth functions:
$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2$, used in most proofs on smooth optimization (Lemma 1.16).

Douglas-Rachford Splitting (DRS) Operator-splitting method for $\min_x f(x) + g(x)$ that alternates reflected proximal steps: $z_{k+1} = \frac{1}{2}(R_{g,t} \circ R_{f,t} + \mathrm{Id})\, z_k$. ADMM is DRS applied to the dual problem (Theorem 3.24).

Duality Gap Difference $p^\star - d^\star \geq 0$ between primal and dual optimal values; zero under strong duality (e.g., Slater's condition). For a feasible pair, $f_0(x) - g(\lambda, \nu) \geq 0$ bounds $f_0(x) - p^\star$ and serves as a stopping test.

Epigraph The set $\{(x, t) : f(x) \leq t\} \subset \mathbb{R}^{n+1}$ lying on or above the graph of f. Convexity of the epigraph is equivalent to convexity of f.

Fenchel Conjugate
$g^*(y) = \sup_x\{\langle y, x\rangle - g(x)\}$. Used in duality theory and primal-dual algorithms (Chambolle-Pock). Satisfies the Fenchel-Young inequality $g(x) + g^*(y) \geq \langle x, y\rangle$.

Firm Non-expansiveness
An operator T is firmly non-expansive if $\|Tx - Ty\|^2 + \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 \leq \|x - y\|^2$. Proximal operators are firmly non-expansive.

FISTA (Fast ISTA)
Accelerated proximal gradient method with rate $O(1/k^2)$ for composite optimization, using Nesterov's momentum; this order matches the lower bound for first-order methods (Sec. 3).

Frank-Wolfe Algorithm
Projection-free method for constrained optimization that replaces the projection step by a Linear Minimization Oracle (LMO). Well suited to nuclear norm and ℓ_1 constraints, whose LMO is cheap.

ISTA (Iterative Shrinkage-Thresholding)
Proximal gradient method for composite optimization $\min f(x) + g(x)$: $x_{k+1} = \mathbf{prox}_{g,t}(x_k - t\nabla f(x_k))$. Converges at rate $O(1/k)$ (Sec. 3).

Lagrange Multiplier
Dual variable λ_i or ν_j associated with an inequality or equality constraint. For a convex problem with strong duality, when the optimal value is differentiable with respect to perturbations of the right-hand sides of the constraints, the optimal multiplier is minus its derivative (admitted) .

Linear Minimization Oracle (LMO)
Subroutine that solves $\mathrm{argmin}_{s\in\mathcal{C}}\langle s, d\rangle$ over the constraint set. For the ℓ_1 ball, the LMO returns a signed standard basis vector; for the nuclear norm ball, a rank-1 matrix from the leading singular vectors.

Moreau's Identity
Decomposition $v = \mathbf{prox}_{g,t}(v) + t\,\mathbf{prox}_{g^*,1/t}(v/t)$, linking the proximal operator of g to that of its conjugate g^*. Useful for computing proximal operators indirectly.

Monotone Operator
A set-valued operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ satisfying $\langle u - v, x - y\rangle \geq 0$ for all $(x, u), (y, v) \in \mathrm{gra}(A)$. Subdifferentials of convex functions and gradients of smooth convex functions are monotone. The optimality condition $0 \in \partial f(x)$ is a monotone inclusion (Sec. 3.5).

Nuclear Norm Sum of singular values $\|X\|_* = \sum_i \sigma_i(X)$. Convex relaxation of the rank function, used in matrix completion and low-rank recovery.

Operator Splitting
Framework for solving $\min f(x) + g(x)$ by handling f and g separately via their proximal operators (resolvents), rather than requiring the full problem structure simultaneously. This amounts to decomposing the monotone inclusion $0 \in (A + B)(x)$ into subproblems involving A and B individually. Includes DRS, ADMM, Chambolle-Pock (Sec. 3.5).

Polyak–Łojasiewicz (PL) Condition
Inequality $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^\star)$, which gives linear convergence of GD for L-smooth f *without requiring convexity*. Implied by strong convexity.

Proximal Operator
$\mathbf{prox}_{g,t}(v) = \operatorname{argmin}_x\{g(x) + \frac{1}{2t}\|x - v\|^2\}$. Generalizes the projection operator; for $g = \lambda\|\cdot\|_1$, it reduces to the soft-thresholding operator.

Reflected Resolvent
$R_{tA} = 2J_{tA} - \mathrm{Id}$. Non-expansive operator central to Douglas-Rachford splitting, where the iteration is driven by composition of two reflected resolvents. For a smooth function g, takes the form $R_{g,t}(u) = x - t\nabla g(x)$ (a gradient step). See Definition 3.15.

Resolvent $J_{tA} = (\mathrm{Id} + tA)^{-1}$. For a maximal monotone operator A, the resolvent is single-valued, defined everywhere (Minty's theorem, admitted) and firmly non-expansive. When $A = \partial g$, it coincides with $\mathbf{prox}_{g,t}$. Zeros of A are fixed points of J_{tA}. See Definition 3.14.

Robbins–Monro Conditions
Classical sufficient conditions on the step sizes for SGD convergence, under suitable assumptions: $\sum t_k = \infty$ and $\sum t_k^2 < \infty$.

SAFE Screening Technique using a duality gap to identify and discard features that have a zero coefficient at the optimum, before or during the optimization (Proposition 2.19).

Slater's Condition
Constraint qualification requiring the existence of a strictly feasible point ($f_i(x) < 0$, $h_j(x) = 0$). Guarantees strong duality for convex problems (Theorem 1.24, admitted).

Strong Convexity
$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2$. For a differentiable f on $\mathbb{R}^n$, it ensures a unique minimizer and gives linear convergence of gradient descent.

Subgradient Generalization of the gradient for non-smooth convex functions: $\xi \in \partial f(x)$ satisfies $f(y) \geq f(x) + \xi^T(y - x)$ for all y. The subdifferential $\partial f(x)$ is a convex set.

Variance Reduction
Class of techniques (SVRG and related methods) that reduce the variance of stochastic gradient estimates with a control variate built from a periodically computed full gradient. Under strong convexity, SVRG converges linearly with $n + 2M$ gradient evaluations of the f_i per epoch (Theorem 2.12, admitted).