We have the following key approximation properties of $\ell(\cdot;x)$ and descent property of (95).

Lemma 1. (a) *For any $x \in \mathrm{dom}P$, we have*

$$f_P(y) \;\geq\; \ell(y;x) \;\geq\; f_P(y) - \frac{L}{2}\|y-x\|^2 \quad \forall y \in \mathrm{dom}P. \tag{97}$$

(b) *For any $x \in \mathrm{dom}P$, letting x_+ solve (95) and we have*

$$\ell(x_+;x) + \frac{L}{2}\|x_+ - x\|^2 \;\leq\; \ell(y;x) + \frac{L}{2}\|y-x\|^2 - \frac{L}{2}\|y-x_+\|^2 \quad \forall y. \tag{98}$$

Proof. (a) Fix any $x \in \mathrm{dom}P$. By convexity of f, we have (see (15))

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle \quad \forall y \in \mathrm{dom}P.$$

Adding $P(y)$ to both sides yields the first inequality in (97). To prove the second inequality, fix $y \in \mathrm{dom}P$ and let $\phi(t) = f(x + t(y-x))$. Then by (92) and the chain rule, ϕ is continuously differentiable on $[0,1]$. By the fundamental theorem of calculus, $\phi(1) - \phi(0) = \int_0^1 \phi'(t)dt$, so that

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y-x\rangle dt \\
&= f(x) + \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y-x\rangle dt \\
&\leq f(x) + \langle \nabla f(x), y-x\rangle + \int_0^1 Lt\|y-x\|^2 dt \\
&= f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2,
\end{aligned}$$

where the inequality uses the Cauchy-Schwarz inequality and the Lipschiz property (92) ($x, y \in \mathrm{dom}P$ and the convexity of $\mathrm{dom}P$ imply $x + t(y-x) \in \mathrm{dom}P$ for $0 \leq t \leq 1$).

(b) For any differentiable convex function $\psi : \mathbb{H} \to \mathbb{R}$, if

$$x_+ = \arg\min_y \psi(y) + P(y),$$

then x_+ is a minimizer of $y \mapsto \langle \psi(x_+), y\rangle + P(y)$, i.e.,

$$\langle \psi(x_+), x_+\rangle + P(x_+) \leq \langle \psi(x_+), y\rangle + P(y) \quad \forall y.$$

(This can be argued by contradiction.) In fact, the converse also holds, though we don't need it. Applying this to $\psi(y) = \langle \nabla f(x), y\rangle + \frac{L}{2}\|y-x\|^2$ (so that $\nabla\psi(y) = \nabla f(x) + L(y-x)$) and rearranging terms yields (98). ∎

Intuitively, Lemma 1 says that $\ell(y;x)$ approximates $f_P(y)$ with second-order error $O(\|y - x\|^2)$ and that the minimizer x_+ of $\ell(y;x)$ yields sufficient descent of second-order. (The second inequality in (97) does not require f to be convex and is useful in extending the method to nonconvex f and proving global convergence. However, the subsequent complexity results hold only for convex f.)

6.2 Proximal gradient method

We saw in the previous subsection examples for which (95) can be solved easily. Given $x \in \mathrm{dom}P$, we solve (95) to obtain a new x_+ and re-iterate. We call this the *proximal gradient method*. When $P \equiv 0$, it reduces to Cauchy's steepest descent method. When P is the indicator function for a closed convex set, it reduces to the gradient projection method of Goldstein, and Levitin and Polyak.[11] To analyze the convergence and iteration complexity of this method, let x^k (or x_k?) denote the new iterate after the kth iteration, so that

$$x^{k+1} \;=\; \arg\min_y \;\ell(y;x^k) + \frac{L}{2}\|y-x^k\|^2, \quad k=0,1,\ldots, \tag{99}$$

with $x^0 \in \mathrm{dom}P$ given. Using Lemma 1, we obtain the following $O(\frac{L}{k})$ bound on the optimality gap after k iterations.

Proposition 5. *For any $x \in \mathrm{dom}P$, we have*

$$f_P(x^k) \le f_P(x) + \frac{L}{k}\frac{\|x-x^0\|^2}{2} \quad \forall k \ge 1,$$

where x^k is given by (99).

Proof. Fix any $x \in \mathrm{dom}P$. For any $k \in \{0,1,\ldots\}$, we have

$$\begin{aligned} f_P(x^{k+1}) &\le \ell(x^{k+1};x^k) + \frac{L}{2}\|x^{k+1}-x^k\|^2 &(100)\\ &\le \ell(x;x^k) + \frac{L}{2}\|x-x^k\|^2 - \frac{L}{2}\|x-x^{k+1}\|^2 \\ &\le f_P(x) + \frac{L}{2}\|x-x^k\|^2 - \frac{L}{2}\|x-x^{k+1}\|^2, &(101)\end{aligned}$$

where the first and third inequalities use Lemma 1(a) and the second inequality uses (99) and Lemma 1(b). Letting $e_k = f_P(x^k) - f_P(x)$ and $\Delta_k = \frac{L}{2}\|x-x^k\|^2$, this simplifies to

$$e_{k+1} \le \Delta_k - \Delta_{k+1}.$$

Thus

$$\Delta_{k+1} \le \Delta_k - e_{k+1} \le \Delta_{k-1} - e_k - e_{k+1} \le \cdots \le \Delta_0 - e_1 - \cdots - e_{k+1}. \tag{102}$$

By (99), the right-hand side of (100) is less than or equal to $\ell(y;x^k) + \frac{L}{2}\|y-x^k\|^2$ for all y. Setting $y = x^k$ yields

$$f_P(x^{k+1}) \le \ell(x^k;x^k) = f_P(x^k).$$

Thus $e_{k+1} \le e_k$ for all k. It follows from (102) that

$$\Delta_{k+1} \le \Delta_0 - (k+1)e_{k+1}.$$

[11] If $P(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{else} \end{cases}$ with $C \subseteq \mathbb{H}$ a nonempty closed convex set, then $x_+ = \arg\min_{y\in C}\langle g,y\rangle + \frac{L}{2}\|y-x\|^2 = \arg\min_{y\in C}\frac{L}{2}\|y-(x-\frac{g}{L})\|^2 = \mathrm{Proj}_C(x-\frac{g}{L})$, where $g = \nabla f(x)$.

Since $\Delta_{k+1} \geq 0$, this implies $(k+1)e_{k+1} \leq \Delta_0$ or $e_{k+1} \leq \frac{\Delta_0}{k+1}$, for all k. Thus $e_k \leq \frac{\Delta_0}{k}$, for all $k \geq 1$. ■

Assuming f_P has a minimizer and taking x^* to be the minimizer nearest to x^0 in Proposition 5, we obtain that $f_P(x^k) \leq f_P(x^*) + \epsilon$ whenever

$$k \geq \frac{L\|x^* - x^0\|^2}{2\epsilon}. \tag{103}$$

Thus the number of iterations to compute an ϵ-optimal solution is $O(\frac{L}{\epsilon})$. If $\inf_x f_P(x)$ is not attained, then we can take x^* to be an $\frac{\epsilon}{2}$-optimal solution. The $O(\frac{1}{\epsilon})$ bound cannot be improved upon as can be seen with the example $\min_{x \leq 0} e^x$. (If f is strongly convex, then this bound can be improved.) The work per iteration is between $O(n)$ and $O(n^3)$ operations for the applications of Section 5. In contrast, the number of iterations for interior-point methods is at best $O(\sqrt{n} \log\left(\frac{\epsilon^0}{\epsilon}\right))$ and the work per iteration is typically between $O(n^3)$ and $O(n^4)$ ops. Thus, for moderate ϵ ($\epsilon = .001$), moderate L (which may depend on n), and large n ($n \geq 10000$), the proximal gradient method can outperform interior-point methods.

In practice, L is typically difficult to estimate or its estimate is too conservative (i.e., too large), leading to slow convergence. A more practical strategy is to start with a guess of L and increase L whenever sufficient descent is not achieved and backtrack:

- Initialize $L > 0$.
- Chek at each iteration k if x^{k+1} satisfies the descent condition (100). If not, then increase L and recompute x^{k+1}.

6.3 Accelerated proximal gradient method

Can the proximal gradient method be accelerated? The steepest descent method is well known to be slowly converging when f is "ill-conditioned." If f is twice differentiable, then one could replace the rough quadratic approximation in (99) with a more accurate quadratic perturbation: $\frac{1}{2}\langle y - x^k, H^k(y - x^k)\rangle$, where $H^k : \mathbb{H} \to \mathbb{H}$ is an approximation of the Hessian of f at x^k. However, the corresponding subproblem will be much more difficult to solve. Alternatively, we can approximate f by not a single linear approximation but by the pointwise maximum of multiple linear approximations. But this too makes the subproblem much more difficult to solve.

Somewhat remarkably, the proximal gradient method can be accelerated "for free" by inserting an extrapolation step in the direction $x^k - x^{k-1}$ at each iteration k. The idea for this goes back to a work of Nesterov in 1983 for the unconstrained case ($P \equiv 0$). The extension to $P \not\equiv 0$ is a very recent result. To motivate this, let's revisit the complexity proof for the proximal gradient method. The key to the proof is the recursion (101). Suppose we can obtain a similar recursion but with L scaled by something tending to zero with k. Then a faster convergence rate would result. How to obtain this? We obviously need to modify (99). One possible modificaction is

$$x^{k+1} \;=\; \arg\min_y \; \ell(y; y^k) + \frac{L}{2}\|y - y^k\|^2, \tag{104}$$

with y^k to be determined. (An alternative would be to replace the second y^k by x^k, but the resulting method and analysis is somewhat more complicated.) Then, as in the proof of Proposition 5, we have that

$$\begin{aligned} f_P(x^{k+1}) &\leq \ell(x^{k+1};y^k) + \frac{L}{2}\|x^{k+1}-y^k\|^2 \\ &\leq \ell(y;x^k) + \frac{L}{2}\|y-y^k\|^2 - \frac{L}{2}\|y-x^{k+1}\|^2 \\ &\leq f_P(y) + \frac{L}{2}\|y-y^k\|^2 - \frac{L}{2}\|y-x^{k+1}\|^2 \quad \forall y, \end{aligned} \tag{105}$$

where the first and third inequalities use Lemma 1(a) and the second inequality uses (104) and Lemma 1(b). So far, nothing new. Now the key step: Recall that we want L to be scaled by something tending to zero. Fix any $x \in \mathrm{dom} P$ and set $y = (1-\theta_k)x^k + \theta_k x$ in the above inequality, with $0 < \theta_k \leq 1$ to be determined. We then factor θ_k out of x to scale L. Specifically, we have

$$\begin{aligned} & f_P(x^{k+1}) \\ \leq\ & f_P((1-\theta_k)x^k + \theta_k x) + \frac{L}{2}\|(1-\theta_k)x^k + \theta_k x - y^k\|^2 - \frac{L}{2}\|(1-\theta_k)x^k + \theta_k x - x^{k+1}\|^2 \\ =\ & f_P((1-\theta_k)x^k + \theta_k x) + \frac{L}{2}\theta_k^2\|x + (\theta_k^{-1}-1)x^k - \theta_k^{-1}y^k\|^2 - \frac{L}{2}\theta_k^2\|x + (\theta_k^{-1}-1)x^k - \theta_k^{-1}x^{k+1}\|^2 \\ =\ & f_P((1-\theta_k)x^k + \theta_k x) + \frac{L}{2}\theta_k^2\|x - z^k\|^2 - \frac{L}{2}\theta_k^2\|x - z^{k+1}\|^2, \end{aligned}$$

where we have rewritten the terms on the next-to-last line to suggest (101) and the last equality is obtained by setting

$$z^k = -(\theta_k^{-1}-1)x^k + \theta_k^{-1}y^k \tag{106}$$

and choosing y^k so that this equals $-(\theta_{k-1}^{-1}-1)x^{k-1} + \theta_{k-1}^{-1}x^k$, which works out to be

$$y^k = x^k + \theta_k(\theta_{k-1}^{-1}-1)(x^k - x^{k-1}). \tag{107}$$

Using also the convexity of f_P, we thus obtain that

$$f_P(x^{k+1}) \leq (1-\theta_k)f_P(x^k) + \theta_k f_P(x) + \frac{L}{2}\theta_k^2\|x-z^k\|^2 - \frac{L}{2}\theta_k^2\|x-z^{k+1}\|^2 \quad \forall k.$$

Letting $e_k = f_P(x^k) - f_P(x)$ and $\Delta_k = \frac{L}{2}\|x-z^k\|^2$, this simplifies to

$$e_{k+1} \leq (1-\theta_k)e_k + \theta_k^2\Delta_k - \theta_k^2\Delta_{k+1}.$$

Dividing both sides by θ_k^2 yields

$$\frac{1}{\theta_k^2}e_{k+1} + \Delta_{k+1} \leq \frac{1-\theta_k}{\theta_k^2}e_k + \Delta_k.$$

This can be rewritten as the recursion

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2}e_{k+1} + \Delta_{k+1} \leq \frac{1-\theta_k}{\theta_k^2}e_k + \Delta_k$$

by choosing θ_{k+1} so that $\frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}$, which works out to be

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}, \tag{108}$$

with $0 < \theta_0 \leq 1$ arbitrary. It is easily seen that $\theta_{k+1} > 0$ and $\frac{\theta_{k+1}}{\theta_k} = \sqrt{1-\theta_{k+1}} < 1$, so that $\theta_{k+1} < \theta_k \leq 1$. Upon propagating the preceding recursion backwards, we have

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} e_{k+1} + \Delta_{k+1} \leq \cdots \leq \frac{1-\theta_0}{\theta_0^2} e_0 + \Delta_0 \quad \forall k.$$

Take $\theta_0 = 1$. Since $\Delta_{k+1} \geq 0$ and $\frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}$, this simplifies to

$$\frac{e_{k+1}}{\theta_k^2} \leq \Delta_0 \quad \forall k.$$

Hence

$$e_k \leq \theta_{k-1}^2 \Delta_0 \quad \forall k \geq 1.$$

Also, an inductive argument shows that $\theta_k \leq \frac{2}{k+2}$ for all k. Moreover, we have from (107) and (106) and taking $\theta_{-1} = 1$ that $z^0 = y^0 = x^0$. We thus arrive at the following improved iteration complexity for the accelerated proximal gradient method.

Proposition 6. *For any $x \in \mathrm{dom} P$, we have*

$$f_P(x^k) \leq f_P(x) + \theta_{k-1}^2 L \frac{\|x - x^0\|^2}{2} \quad \forall k \geq 1,$$

where x^k is given by (104)*, with y^k given by* (107)*, θ_k given by* (108)*, and $\theta_{-1} = \theta_0 = 1$. Moreover, $\theta_{k-1} \leq \frac{2}{k+1}$ for all $k \geq 1$.*

Assuming f_P has a minimizer and taking x^* to be the minimizer nearest to x^0 in Proposition 6, we obtain that $f_P(x^k) \leq f_P(x^*) + \epsilon$ whenever $\frac{2}{(k+1)^2} L \|x^* - x^0\|^2 \leq \epsilon$ or, equivalently,

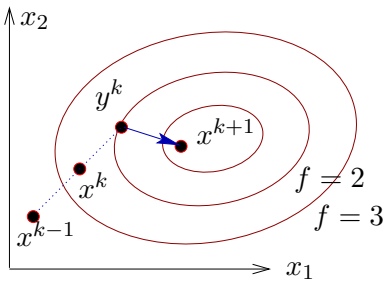
$$k \geq \sqrt{\frac{2L\|x^* - x^0\|^2}{2\epsilon}} - 1. \tag{109}$$

Thus the number of iterations to compute an ϵ-optimal solution is $O(\sqrt{\frac{L}{\epsilon}})$. This significantly improves on $O(\frac{L}{\epsilon})$ for the proximal gradient method and the only additional work per iteration is the computation of y^k, which takes only $O(n)$ ops for $\mathbb{H} = \mathbb{R}^n$ and $O(n^2)$ ops for $\mathbb{H} = \mathbb{S}^n$. In practice, the accelerated method seems invariably faster. Here, L can be similarly adjusted using backtracking, but with (105) replacing (100).

As $k \to \infty$, we have $\theta_k \to 0$ and $\frac{\theta_k}{\theta_{k-1}} = \sqrt{1-\theta_k} \to 1$, so that (107) yields

$$y^k \approx x^k + (x^k - x^{k-1}).$$

Thus y^k is asymptotically an isometric extrapolation from x^{k-1} towards x^k. The accelerated method takes a gradient step from the extrapolated point y^k instead of x^k.



How to terminate the proximal gradient method (99) or (104)? We can use (103) and (109), but this requires estimating $\|x^* - x^0\|$ and can be too conservative. However, if f can be expressed in the saddle form

$$f(x) = \max_{v \in V} \phi(x, v),$$

for some suitable function ϕ, then we can accelerate termination of the method using duality gap. Introducing the Lagrangian

$$L(x, v) := \phi(x, v) + P(v),$$

the primal problem (91) corresponds to $\min_x \max_v L(x, v)$ and the dual problem corresponds to $\max_v \min_x L(x, v)$. Letting $d_P(v) := \min_x L(x, v)$ ("dual function"), we compute (maybe every 5 or 10 iterations) a candidate dual solution

$$v^k \;=\; \arg\max_v \phi(x^k, v)$$

and check that $f_P(x^k) - d_P(v^k) \le \epsilon$. In fact, assuming furthermore that $\mathrm{dom} P$ is bounded (such as the dual image denoising problem (87)), it can be shown that

$$0 \;\le\; f_P(x^{k+1}) - q_P(\bar{v}^k) \;\le\; \theta_k^2 \frac{L}{2} \max_{x \in \mathrm{dom} P} \|x - x^0\|^2 \quad \forall k \ge 0.$$

where x^{k+1}, y^k, θ_k are given by (104), (107), (108), with $\theta_{-1} = \theta_0 = 1$, and we let

$$\hat{v}^k \;=\; \arg\max_v \phi(y^k, v), \qquad \bar{v}^k \;=\; (1 - \theta_k)\bar{v}^{k-1} + \theta_k \hat{v}^k.$$

with $\bar{v}^{-1} = 0$.

In some applications, P is the indicator function for the unit simplex (or a Cartesian product of unit simplices) in $\mathbb{H} = \mathbb{R}^n$. An example is when x is a probability distribution. Then (95) reduces to a projection onto the unit simplex, which can be computed in $O(n)$ ops. However, the algorithms to compute this projection tend to be complicated (e.g., bisection of breakpoints using a linear-time median-finding algorithm). An alternative is to replace the quadratic proximity term $\|y - x\|^2/2$ in (95) by a proximity term of the form

$$D(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle,$$

where the kernel function h is strongly convex on $\mathrm{dom} P$ and differentiable on $\mathrm{dom} P \cap (0, \infty)^n$. ($D$ was first used in a 1965 paper of Bregman.) For $h(x) = \|x\|^2/2$, we recover $D(y, x) = \|y - x\|^2/2$. For $h(x) = \sum_i x_i \log x_i$, $D(y, x)$ is the so-called Kullback-Liebler divergence for probability distributions and, more importantly, (95) has a closed-form solution computable in

$O(n)$ ops. For this D, an accelerated proximal gradient method can be developed with $O(\frac{1}{k^2})$ convergence rate.

Can the convergence rate be further improved? Reviewing the above proof, we see that the convergence rate can be improved to $O(\frac{1}{k^p})$ $(p > 2)$ if we can replace $\|\cdot\|^2$ in the proof by $\|\cdot\|^p$. However, this may require using a higher-order approximation of f in $\ell(\cdot; x)$ (see Lemma 1(a)), which means significantly more work to solve (95). Thus the improvement would not come "for free".

6.4 Proximal minimization method

In the proximal gradient method of the last two subsections, we linearize f at each iteration so that the resulting subproblem is simpler. What if we do not linearize? In other words, we generate

$$x^{k+1} \;=\; \arg\min_y \; f_P(y) + \frac{\tau_k}{2}\|y - x^k\|^2, \quad k = 0, 1, \ldots, \tag{110}$$

with $\tau_k > 0$ and $x^0 \in \mathrm{dom} P$. On first appearance, this seems like a silly idea, since we seem to be making the problem harder by adding a quadratic to the objective function and then we solve this harder problem multiple times. However, the dual of (110) has nice structure and can be solved efficiently (inexactly) by a Newton-type method. In fact, this approach is not unlike the interior-point method, whereby we added a log-barrier to the objective function and solved the resulting subproblem inexactly using Newton's method.

We call (110), first studied by Martinet in the 70's, the proximal minimization method. It was shown by Terry Rockafellar that this method is equivalent to the augmented Lagrangian method of Hesteness and Powell, applied to the dual problem. Very recently, this method was shown to be effective (faster than existing methods, including interior-point method) for solving large SDP with $m = \Theta(n^2)$ constraints. Specifically, consider the conic optimization problem (18) and its dual (19). Applying the proximal minimization method (110) to the primal problem (18) yields

$$x^{k+1} \;=\; \arg\max_{y \in K, \mathcal{A}y = b} \langle c, y\rangle - \frac{\tau_k}{2}\|y - x^k\|^2.$$

By introducing the primal augmented Lagrangian

$$L_k(y, \mu) = \langle c, y\rangle - \frac{\tau_k}{2}\|y - x^k\|^2 + (b - \mathcal{A}y)^T\mu,$$

the above subproblem corresponds to $\max\limits_{y \in K} \min\limits_{\mu \in \mathbb{R}^m} L_k(y, \mu)$. Its dual corresponds to $\min\limits_{\mu \in \mathbb{R}^m} \max\limits_{y \in K} L_k(y, \mu)$, which works out to be (also using the fact that $\langle \mathrm{Proj}_K(u), \mathrm{Proj}_{K^\circ}(u)\rangle = 0$ for any $u \in \mathbb{H}$)

$$\min_{\mu \in \mathbb{R}^m} \frac{\tau_k}{2}\left(\left\|\mathrm{Proj}_K(x^k + \frac{c - \mathcal{A}^*\mu}{\tau_k})\right\|^2 - \|x^k\|^2\right) + b^T\mu. \tag{111}$$

Moreover, it can be shown that

$$x^{k+1} = \mathrm{Proj}_K(x^k + \frac{c - \mathcal{A}^*\mu^k}{\tau_k}), \tag{112}$$

where μ^k solves the dual augmented Lagrangian subproblem (111). It can be shown that the objective function in (111) is convex, continuously differentiable, but not twice differentiable. Nonethelss, for SDP, it is possible to solve (111) efficiently using a Newton-type method, with each Newton direction computed inexactly using a preconditioned conjugate gradient (PCG) method. Unlike interior-point method, the linear equation solved by the PCG method does not become progressively more ill-conditioned, which is a key advantage.

6.5 Proximal cutting-plane method

Somewhere between the proximal gradient methods of Subsections 6.2 and 6.3, which approximate f by a single linear function at each iteration, and the proximal minimization method of Subsection 6.4, which uses the full f at each iteration, is a method that approximates f by the pointwise-maximum of a finite number of linear functions. We call this the proximal cutting-plane method (alternatively, "bundle method"), as the graph of each linear function may be viewed as a plane in $\mathbb{H} \times \mathbb{R}$ supporting epi$f$. The basic version of this method has the form

$$x^{k+1} \;=\; \arg\min_y \; \max_{x \in \mathcal{X}^k} \ell(y, x) + \frac{\tau_k}{2}\|y - x^k\|^2, \quad k = 0, 1, \ldots, \tag{113}$$

with $\tau_k > 0$, $\mathcal{X}^k \subset \mathrm{dom} P$, and $x^0 \in \mathrm{dom} P$. Typically $\mathcal{X}^k \subset \{x^0, x^1, \cdots, x^k\}$. The work per iteration is generally much greater than for the proximal gradient method, but less than for the proximal minimization method. Iteration complexity? Accelerated methods?