ELE 538: Large-Scale Optimization
Homework 2

Zachary Hervieux-Moore

Monday 26th March, 2018

Exercise 1: Conjugate subgradient theorem: Suppose f is convex. Show that the following two statements are equivalent.

i) $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = f(\boldsymbol{x}) + f^*(\boldsymbol{y})$

ii) $\boldsymbol{y} \in \partial f(\boldsymbol{x})$

Remark: this also means that the above statements are equivalent to $\boldsymbol{x} \in \partial f^*(\boldsymbol{y})$.

Answer: Starting from i),

$$
\begin{aligned}
\langle \boldsymbol{x}, \boldsymbol{y} \rangle &= f(\boldsymbol{x}) + f^*(\boldsymbol{y}) \\
&= f(\boldsymbol{x}) + \sup_{z} \{\langle \boldsymbol{z}, \boldsymbol{y} \rangle - f(\boldsymbol{z})\}
\end{aligned}
$$

$$
\begin{aligned}
&\iff \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq f(\boldsymbol{x}) + \langle \boldsymbol{z}, \boldsymbol{y} \rangle - f(\boldsymbol{z}) \quad \forall \boldsymbol{z} \\
&\iff f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{z} - \boldsymbol{x}, \boldsymbol{y} \rangle \quad \forall \boldsymbol{z} \\
&\iff \boldsymbol{y} \in \partial f(\boldsymbol{x})
\end{aligned}
$$

Exercise 2: Alternating projections for LP feasibility: We consider the problem of finding a point $\boldsymbol{x} \in \mathbb{R}^n$ that satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \succeq 0$, where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, with $m < n$.

a) Work out alternating projections for this problem. (In other words, explain how to compute (Euclidean) projections onto $\{\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ and $\mathbb{R}^n_+$.)

b) Implement your method, and try it on one or more problem instances with $m = 500$, $n = 2000$. With $\boldsymbol{x}^k$ denoting the k^{th} iterate after projection onto $\mathbb{R}^n_+$, plot $\|\boldsymbol{A}\boldsymbol{x}^k - \boldsymbol{b}\|_2$, the residual of the equality constraint. (This should converge to zero; you can terminate when this norm is smaller than 10^{-5}.)

c) A general method that can speed up alternating projections is to over-project, which means replacing the simple projection $\boldsymbol{x}^+ = \mathcal{P}(\boldsymbol{x})$ with $\boldsymbol{x}^+ = \boldsymbol{x} + \gamma(\mathcal{P}(\boldsymbol{x}) - \boldsymbol{x})$, where $\gamma \in [1, 2)$. (When $\gamma = 1$, this reduces to standard projection.) It is not hard to show that alternating projections, with over-projection, converges to a point in the intersection of the sets. Implement over-projection and experiment with the over-projection factor γ, observing the effect on the number of iterations required for convergence.

Answer:

a) The projection onto $\{\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ from a point $\boldsymbol{z}$ is solved by the following minimization problem

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 \\ \text{s.t.} \ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

This is a convex problem and so it is well behaved so we just find the KKT conditions

$$\begin{aligned} \boldsymbol{x} - \boldsymbol{z} + \boldsymbol{A}^T\lambda = 0 \\ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

Manipulating the first equation

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{z} + \boldsymbol{A}\boldsymbol{A}^T\lambda = 0 \\ \implies \lambda = (\boldsymbol{A}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}) \end{aligned}$$

Which gives us the solution

$$\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{z})$$

Then, projecting onto $\mathbb{R}^n_+$ is simply making the negative entries equal to 0

$$\mathcal{P}_{\mathbb{R}^n_+}(\boldsymbol{x})_i = \max(0, x_i) \quad \forall i$$

b) The code used to generate the figure is attached below. Figure 1 shows the convergence to a stationary point. However, due to round off errors with the large dimension size, Matlab does not converge to 0. It does for smaller problems ($m = 5, n = 2000$).

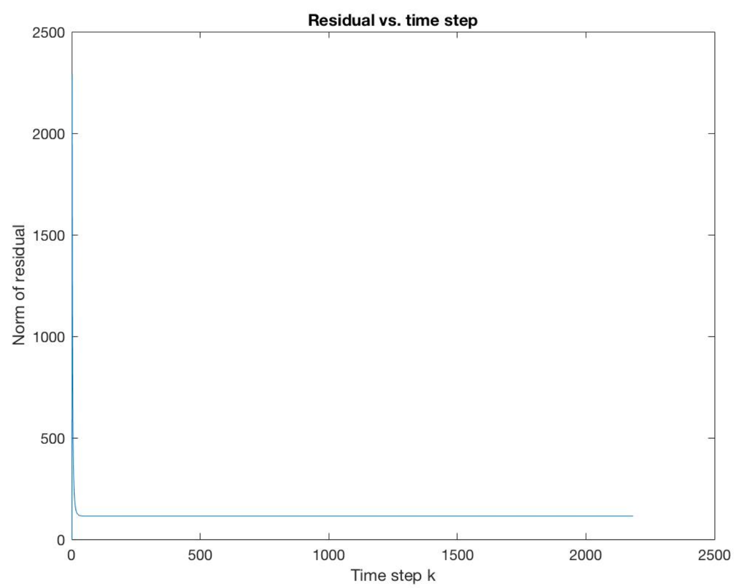

Figure 1: Plot of $\|\boldsymbol{A}\boldsymbol{x}^k - \boldsymbol{b}\|$

Code Appendix:

```
clear;
clc;

m = 500;
n = 2000;
```

```
A = rand(m,n);
b = rand(m,1);
x(:,1) = rand(n,1);
i = 1;

while (norm(A*x(:,i) - b,2) > 10^-5)
    i = i + 1;
    % project into affine set
    x(:,i) = x(:,i-1) + A'/(A*A')*(b-A*x(:,i-1));

    i = i + 1;
    % project into positive orthant
    x(:,i) = max(0, x(:,i-1));

    res(ceil(i/2)) = norm(A*x(:,i) - b,2);
end

plot(res)
```

c) The code used to generate the figure is attached below. Notice that I used a smaller setting so that I got convergence so I could compare different λ. The figure below shows that convergence does speed up with an increased λ but starts to slow down as λ approaches 2.

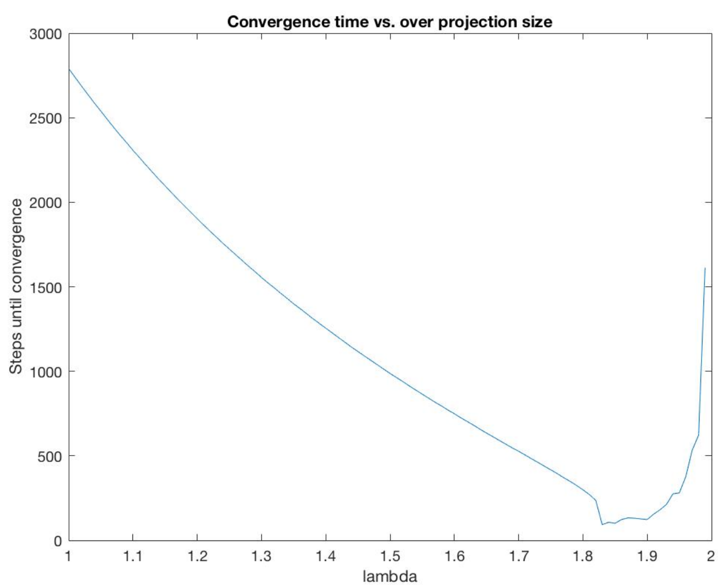

Figure 2: Plot of convergence time vs. λ

Code Appendix:

```
clear;
clc;

m = 5;
n = 2000;

A = rand(m,n);
b = rand(m,1);
x(:,1) = zeros(n,1);
lambda = zeros(1,100);
j = 1;

for gamma=1:0.01:1.99
    clear x;
    x(:,1) = zeros(n,1);
    i = 1;
    while (norm(A*x(:,i) - b,2) > 10^-5)
        i = i + 1;
        % project into affine set
        x(:,i) = x(:,i-1) + gamma*(x(:,i-1) + A'/(A*A')*(b-A*x(:,i-1))
            -x(:,i-1));

        i = i + 1;
        % project into positive orthant
        x(:,i) = x(:,i-1) + gamma*(max(0, x(:,i-1))-x(:,i-1));
    end
    lambda(j) = i;
    j = j + 1;
end

plot(1:0.01:1.99, lambda)
```

Exercise 3: Minimizing expected Bregman divergence: Let $\boldsymbol{z}$ be a random vector with distribution $\mathbb{P}$, and consider the following optimization problem

$$\text{minimize}_{\boldsymbol{x}} \ \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[D_{\varphi}(\boldsymbol{z}, \boldsymbol{x})]$$

for some strongly convex φ. Find the minimizer of this problem.

Answer: We substitute the definition of Bregman divergence into the optimizaiton problem and check the first order conditions for an optimal point.

$$\begin{aligned}
&\text{minimize}_{\boldsymbol{x}} \ \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[D_{\varphi}(\boldsymbol{z}, \boldsymbol{x})] \\
&= \text{minimize}_{\boldsymbol{x}} \ \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\varphi(\boldsymbol{z}) - \varphi(\boldsymbol{x}) - \langle \nabla\varphi(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}\rangle] \\
&= \text{minimize}_{\boldsymbol{x}} \ \langle \nabla\varphi(\boldsymbol{x}), \boldsymbol{x} - \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]\rangle - \varphi(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\varphi(\boldsymbol{z})]
\end{aligned}$$

Where we used the linearity of expectation to bring it into the inner product. Now, we take the gradient of the above with respect to $\boldsymbol{x}$.

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \implies & \nabla\varphi(\boldsymbol{x}) + \langle \nabla^2\varphi(\boldsymbol{x}), \boldsymbol{x} - \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]\rangle - \nabla\varphi(\boldsymbol{x}) \\
&= \langle \nabla^2\varphi(\boldsymbol{x}), \boldsymbol{x} - \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]\rangle
\end{aligned}$$

Of course, the above is equal to 0 if $\boldsymbol{x} = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]$. Now, we must show that this is indeed a minimum. We need to check the second order conditions. So we differentiate with respect to $\boldsymbol{x}$ again and evaluate at $\boldsymbol{x} = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]$.

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \implies & \langle \nabla^3\varphi(\boldsymbol{x}), \boldsymbol{x} - \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]\rangle + \nabla^2\varphi(\boldsymbol{x}) \\
\boldsymbol{x} = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}] \implies & \nabla^2\varphi(\boldsymbol{x})
\end{aligned}$$

As $\varphi(\cdot)$ is strongly convex, we have that this is PSD and hence $\boldsymbol{x} = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{P}}[\boldsymbol{z}]$ is indeed a minimum.

Exercise 4: Exponentiated gradient:

a) Consider the mirror descent update rule with KL divergence

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle + \frac{1}{\eta_t}\mathrm{KL}(\boldsymbol{x}\|\boldsymbol{x}^t) \right\}$$

where $KL(\boldsymbol{x}\|\boldsymbol{z}) := \sum_i x_i \log \frac{x_i}{z_i}$ and $\mathcal{C} = \Delta := \{\boldsymbol{x} \in \mathbb{R}^n_+ | \sum_{i=1}^n x_i = 1\}$. Show that if $\boldsymbol{x}^t \in \Delta$, then

$$x_i^{t+1} = \frac{x_i^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_i)}{\sum_{j=1}^n x_j^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_j)}, \quad 1 \le i \le n$$

b) Consider the mirror descent update rule

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle + \frac{1}{\eta_t} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^t) \right\}$$

where $D_\varphi(\boldsymbol{x}, \boldsymbol{z}) := \sum_i x_i \log \frac{x_i}{z_i} - x_i + z_i$ is the generalized KL divergence and $\mathcal{C} = \mathbb{R}^n_+$ is the positive orthant. When $\boldsymbol{x}^t \in \mathcal{C}$, find a closed-form expression for the mirror descent update.

Answer:

a) As the function is differentiable, we simply differentiate and set to 0.

$$\begin{aligned}
[\nabla_{\boldsymbol{x}}]_i = [\nabla f(\boldsymbol{x}^t)]_i + \frac{1}{\eta_t}\left(\log \frac{x_i}{x_i^t} + 1\right) = 0 \\
\implies \log \frac{x_i}{x_i^t} + 1 = -\eta_t[\nabla f(\boldsymbol{x}^t)]_i \\
\implies x_i = \frac{x_i^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_i)}{e}
\end{aligned}$$

It is simple to check that this is indeed a minimum by checking the second order condition. Since $\boldsymbol{x}^t \in \Delta$, we have that all the components above are positive. However, we must normalize so that $\boldsymbol{x} \in \Delta$. Normalize by the $\|\cdot\|_1$ norm is justified as we are projecting onto Δ. Thus,

$$x_i^{t+1} = \frac{x_i}{\|\boldsymbol{x}\|_1} = \frac{x_i^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_i)}{\sum_{j=1}^n \boldsymbol{x}_j^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_j)}$$

b) We repeat the procedure above

$$[\nabla_{\boldsymbol{x}}]_i = [\nabla f(\boldsymbol{x}^t)]_i + \frac{1}{\eta_t}\left(\log \frac{x_i}{x_i^t} + 1 - 1\right) = 0$$
$$\implies x_i = x_i^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_i)$$

Thus, we have $x_i^{t_1} = x_i^t \exp(-\eta_t[\nabla f(\boldsymbol{x}^t)]_i)$ which is in the positive orthant as x_i^t is aswell.

Exercise 5: Proximal operators:

a) Suppose $f(\boldsymbol{x}) = \sum_{i=1}^{n} w_i|x_i|$ with $w_i \geq 0$. Computer $\mathrm{prox}_f(\boldsymbol{x})$.

b) Show that if $f(\boldsymbol{x}) = g(a\boldsymbol{x} + \boldsymbol{b})$ with $a \neq 0$, then

$$\mathrm{prox}_f(\boldsymbol{x}) = \frac{1}{a}(\mathrm{prox}_{a^2 g}(a\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b})$$

c) Show that if $f(\boldsymbol{x}) = g(\boldsymbol{Q}\boldsymbol{x})$ with $\boldsymbol{Q}$ orthogonal (i.e. $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$), then

$$\mathrm{prox}_f(\boldsymbol{x}) = \boldsymbol{Q}^T \mathrm{prox}_g(\boldsymbol{Q}\boldsymbol{x})$$

d) Let $f(\boldsymbol{x}) = x_{[1]} + \cdots + x_{[k]}$, where $x_{[i]}$ is the i^{th} largest entry of $\boldsymbol{x}$. Compute $\mathrm{prox}_f(\boldsymbol{x})$.

Answer:

a) We have

$$\mathrm{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + \sum_{i=1}^{n} w_i|z_i| \right\}$$

First, we handle the differentiable components. Again, we check first order and second order conditions. One yields

$$z_i = x_i - w_i \text{ if } x_i > w_i$$
$$z_i = x_i + w_i \text{ if } x_i < -w_i$$

Finally, between $-w_i \leq x_i \leq w_i$, we have that the quadratic term is smaller than the absolute term, so we pick $z_i = 0$ which defines the proximal operator for the 3 different cases.

b) We have

$$\mathrm{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + g(a\boldsymbol{z} + b) \right\}$$

Manipulating the above, multiplying by a positive scalar (a^2) keeps the arg min unchanged,

$$\begin{aligned}
&= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|a\boldsymbol{z} + \boldsymbol{b} - a\boldsymbol{x} - \boldsymbol{b}\|^2 + a^2 g(a\boldsymbol{z} + \boldsymbol{b}) \right\} \\
&= \arg\min_{\boldsymbol{z}'} \left\{ \frac{1}{2} \|\boldsymbol{z}' - a\boldsymbol{x} - \boldsymbol{b}\|^2 + a^2 g(\boldsymbol{z}') \right\}
\end{aligned}$$

Where we made the substitute $a\boldsymbol{z} + \boldsymbol{b} = \boldsymbol{z}'$. This is valid since the arg min is still over all $z' \in \mathbb{R}^n$. Thus, we have that $a\boldsymbol{z}^* + \boldsymbol{b} = \boldsymbol{z}'^*$ or $\text{prox}_f(\boldsymbol{x}) = \frac{1}{a}(\text{prox}_{a^2 g}(a\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b})$.

c) We have

$$\begin{aligned}
\text{prox}_f(\boldsymbol{x}) &= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{x}\|^2 + g(\boldsymbol{Q}\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{z} - \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{x}\|^2 + g(\boldsymbol{Q}\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{z}'} \left\{ \frac{1}{2} \|\boldsymbol{Q}^T\boldsymbol{z}' - \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{x}\|^2 + g(\boldsymbol{z}') \right\}
\end{aligned}$$

Where the substitution $\boldsymbol{z}' = \boldsymbol{Q}\boldsymbol{z}$ is justified as $\boldsymbol{z}'$ still spans $\mathbb{R}^n$ by the orthogonality of $\boldsymbol{Q}$. Now, we also note by the orthogonality of $\boldsymbol{Q}$ that

$$\|\boldsymbol{Q}^T\boldsymbol{z}' - \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{x}\|^2 = \|\boldsymbol{z}' - \boldsymbol{Q}\boldsymbol{x}\|^2$$

Thus we have

$$= \arg\min_{\boldsymbol{z}'} \left\{ \|\boldsymbol{z}' - \boldsymbol{Q}\boldsymbol{x}\|^2 + g(\boldsymbol{z}') \right\}$$

That is, $\boldsymbol{z}'^* = \boldsymbol{Q}\boldsymbol{z}^*$, or $\text{prox}_f(\boldsymbol{x}) = \boldsymbol{Q}^T \text{prox}_g(\boldsymbol{Q}\boldsymbol{x})$, as $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$.

d) We can rewrite $f(\boldsymbol{x})$ as $f(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in \mathcal{C}} \boldsymbol{y}^T\boldsymbol{x}$ where $\mathcal{C} = \{\boldsymbol{y} : 0 \preceq \boldsymbol{y} \preceq 1, 1^T\boldsymbol{y} = k\}$. That is, the solution to this optimization function is the k largest entries of $\boldsymbol{x}$. Now we note that

$$f^*(\boldsymbol{x}) = \delta_{\mathcal{C}}(\boldsymbol{x})$$

That is, the Fenchel conjugate of $f(\cdot)$ is the indicator function of $\mathcal{C}$. Now using Moreau decomposition, we have

$$\begin{aligned}
\operatorname{prox}_f(\boldsymbol{x}) &= \boldsymbol{x} - \operatorname{prox}_{f^*}(\boldsymbol{x}) \\
&= \boldsymbol{x} - \arg\min_z \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 + \delta_{\mathcal{C}}(\boldsymbol{z}) \right\} \\
&= \boldsymbol{x} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})
\end{aligned}$$

That is, all one needs to do is compute the projection of $\boldsymbol{x}$ onto $\mathcal{C}$. Which is a projection onto a polyhedral set which is similar (different form) as to what was done in problem 2. The exact solution is easily derived using Lagrange multipliers.

Exercise 6: Extended Moreau decomposition: Let f be closed and convex. Show that for any $\lambda > 0$ and any $\boldsymbol{x}$, one has

$$\boldsymbol{x} = \mathrm{prox}_{\lambda f}(\boldsymbol{x}) + \lambda \mathrm{prox}_{\frac{1}{\lambda} f^*}(\boldsymbol{x}/\lambda)$$

Answer: We prove this by applying Moreau decomposition to λf. Recall that Moreau decomposition is

$$\boldsymbol{x} = \mathrm{prox}_f(\boldsymbol{x}) + \mathrm{prox}_{f^*}(\boldsymbol{x})$$

So, we have to compute the conjugate of λf to get the Extended Moreau decomposition. So we have

$$(\lambda f)^*(\boldsymbol{x}) = \sup_{\boldsymbol{z}} \{\langle \boldsymbol{z}, \boldsymbol{x} \rangle - (\lambda f)(\boldsymbol{z})\}$$

Multiplying and dividing by λ yields, we get that the above is equivalent to

$$\begin{aligned} &= \lambda \sup_{\boldsymbol{z}} \{\langle \boldsymbol{z}, \boldsymbol{x}/\lambda \rangle - (f)(\boldsymbol{z})\} \\ &= \lambda f^*(\boldsymbol{x}/\lambda) \end{aligned}$$

Thus we have that $(\lambda f)^*(\boldsymbol{x}) = \lambda f^*(\boldsymbol{x}/\lambda)$ which we now plug into our proximal operator and use the fact proved in 5b) $(g = f^*, a = 1/\lambda, \boldsymbol{b} = 0)$ to get

$$\begin{aligned} \mathrm{prox}_{(\lambda f)^*}(\boldsymbol{x}) &= \mathrm{prox}_{\lambda g}(\boldsymbol{x}) \\ &= \lambda \mathrm{prox}_{\frac{1}{\lambda^2} \cdot \lambda g}(\boldsymbol{x}/\lambda) \\ &= \lambda \mathrm{prox}_{\frac{1}{\lambda} f^*}(\boldsymbol{x}/\lambda) \end{aligned}$$

Which proves the claim that

$$\boldsymbol{x} = \mathrm{prox}_{\lambda f}(\boldsymbol{x}) + \lambda \mathrm{prox}_{\frac{1}{\lambda} f^*}(\boldsymbol{x}/\lambda)$$