SDS 385 Exercise Set

Kevin Song

October 19, 2016

The Proximal Operator

(A)

The proximal operator for the linear approximation of f at x_0 can be written and reduced as follows:

$$\begin{aligned}
\underset{\gamma}{\text{prox}}\, \hat{f}(x; x_0) &= \underset{\gamma}{\text{prox}} \left[f(x_0) + (x - x_0)^T \nabla f(x_0) \right] \\
&= \arg\min_z \left[f(x_0) + (z - x_0)^T \nabla f(x_0) + \frac{1}{2\gamma} \|z - x\|_2^2 \right] \\
&= \arg\min_z \left[(z - x_0)^T \nabla f(x_0) + \frac{1}{2\gamma} \|z - x\|_2^2 \right] \\
&= \arg\min_z \left[z^T \nabla f(x_0) - x_0^T \nabla f(x_0) + \frac{1}{2\gamma} \left(z^T z - 2z^T + x^T x \right) \right] \\
&= \arg\min_z \left[z^T \nabla f(x_0) + \frac{1}{2\gamma} \left(z^T z - 2z^T x \right) \right]
\end{aligned}$$

where, in the last step, I have thrown away any terms that are constant with respect to z. Since we are taking the argmin over z, these terms are unimportant—alternatively, since we need to take the gradient with respect to z to solve this minimization, these terms will be zero anyways.

To solve for this argmin, we take the gradient of this expression w.r.t z and set the result equal to zero:

$$\begin{aligned}
&\nabla_z \left[z^T \nabla f(x_0) + \frac{1}{2\gamma} \left(z^T z - 2z^T x \right) \right] \\
&= \nabla f(x_0) + \frac{1}{\gamma}(z - x) = 0 \\
&\implies z = x - \gamma \nabla f(x_0)
\end{aligned}$$

This shows that the solution to the proximal operator of the linear approximation of the function is the gradient descent step.

(B)

Consider a log-likelihood of the form $\ell(x) = \frac{1}{2} x^T P x - q^T x + r$. The proximal operator of this function, with parameter $\frac{1}{\gamma}$ is

$$\begin{aligned}
\operatorname*{prox}_{\frac{1}{\gamma}} \ell(x) &= \operatorname*{prox}_{\frac{1}{\gamma}} \frac{1}{2}x^T Px - q^T x + r \\
&= \arg\min_z \frac{1}{2}z^T Pz - q^T z + r + \frac{\gamma}{2}(z-x)^T(z-x) \\
&= \arg\min_z \frac{1}{2}\left(z^T Pz + \gamma z^T Iz\right) - (q^T z + \gamma x^T z) + r + x^T x \\
&= \arg\min_z \frac{1}{2}z^T (P+\gamma I)z - (q+\gamma x)^T z + r + x^T x
\end{aligned}$$

We know that the minimum of the quadratic form $\frac{1}{2}z^T Az + b^T z + c$ is given by the solution to $Az - b = 0$ or $z = A^{-1}b$, so the minimum to this likelihood is

$$\operatorname*{prox}_{\frac{1}{\gamma}} \ell(x) = (P+\gamma I)^{-1}(\gamma x + q)$$

In part B, the likelihood of such a sample is (PDF from Wikipedia because I'm lazy)

$$\mathcal{L}(y_1,\ldots,y_n;x) = \frac{1}{\left(\sqrt{2\pi}\right)^k\sqrt{\det\Omega^{-1}}}\exp\left(-\frac{1}{2}(y-Ax)^T\Omega(y-Ax)\right)$$

Taking the log of both sides, we find that

$$\begin{aligned}
\log\mathcal{L}(y_1,\ldots,y_n;x) &= \log\left[\frac{1}{\sqrt{2\pi}^k\sqrt{\det\Omega^{-1}}}\exp\left(-\frac{1}{2}(y-Ax)^T\Omega(y-Ax)\right)\right] \\
&= \log\frac{1}{\sqrt{2\pi}^k\sqrt{\det\Omega^{-1}}} + \log\exp\left(-\frac{1}{2}(y-Ax)^T\Omega(y-Ax)\right) \\
&= -\left(\log\sqrt{2\pi}^k + \log\sqrt{\det\Omega^{-1}} + \frac{1}{2}(y-Ax)^T\Omega(y-Ax)\right) \\
&= -C - \frac{1}{2}\left(x^T A^T\Omega Ax - 2y^T\Omega Ax + y^T\Omega y\right)
\end{aligned}$$

This shows us that

$$-\log\mathcal{L} = \frac{1}{2}\log\det\Omega^{-1} + \frac{1}{2}\log\sqrt{2\pi}^k + \frac{1}{2}\left(x^T A^T\Omega Ax - 2y^T\Omega Ax + y^T\Omega y\right)$$

This can be fit into the quadratic form by choosing $P = A^T\Omega A$, $q = y^T\Omega A$, and $r = \frac{1}{2}\left(\log\det\Omega^{-1} + \log\sqrt{2\pi}^k + y^T\Omega y\right)$.

(C)

$$\operatorname*{prox}_{\gamma}\phi(x) = \arg\min_z \tau\|z\|_1 + \frac{1}{2\gamma}\|z-x\|_2^2 = \arg\min_z \tau\sum_i |z_i| + \frac{1}{2\gamma}\sum_i (z_i - x_i)^2$$

Since the z_i are independent of each other, by minimizing each element of the summation, we minimize the overall sum (this would not be true if z were constrained in some form).

From Exercise 5, we know that the solution to

$$\arg\min_{z_i} \tau|z_i| + \frac{1}{2\gamma}(z_i - x_i)^2$$

is given by

$$\operatorname{sign}(x_i)(|x_i| - \gamma\tau)_+$$

Proximal Gradient Method

(A)

To prove this, we show that the provided form produces the correct solution:

$$\begin{aligned}
\operatorname*{prox}_{\gamma} \phi(u) &= \arg\min_z \phi(z) - \frac{1}{2\gamma}\|z - x_0 + \gamma\nabla\ell(x_0)\|_2^2 \\
&= \arg\min_z \phi(z) - \frac{1}{2\gamma}[z^Tz - z^Tx_0 + z^T\gamma\nabla\ell(x_0)x_0^Tz + x_0^Tx_0 - x_0^T\gamma\nabla\ell(x_0) \\
&\quad + \gamma\nabla\ell(x_0)^Tz + \gamma\nabla\ell(x_0)^Tx_0 + \gamma^2\nabla\ell(x_0)^T\nabla\ell(x_0)]
\end{aligned}$$

where the second step involves explicitly multiplying out the norm. We can now rearrange this mess and remove some constant terms to reveal that

$$\begin{aligned}
&\arg\min_z \left[\phi(z) - \frac{1}{2\gamma}\left([z^Tz - 2z^Tx_0 + x_0^Tx_0] + 2\gamma z^T\nabla\ell(x_0)\right)\right] \\
&= \arg\min_z \left[\phi(z) - \frac{1}{2\gamma}\|z - x\|_2^2 + z^T\nabla\ell(x_0)\right]
\end{aligned}$$

which is the exact same minimization problem as minimizing $\widetilde{f}$ (up to alpha equivalence, replacing the name z with x).

(B)

```
function calc_gradient(X,y,β){
    // We already know how to do this, omitted for brevity
}

function calc_gamma(){
  return 0.42 //Not going to worry about gamma yet. Just use a constant
}

function solve_prox(x, γ){
  for each $x_i${
    $z_i$ = sign($x_i$) * max((abs($x_i$) - γ), 0)
  }
}

function proximal_gradient(X,y,β){
  repeat until convergence{
     grad = calc_gradient(X,y,β)
     γ = calc_gamma()
     u = β − γ*grad
     β = solve_prox(u, γ)
  }
  }
```

The big costs in this function are calculating the gradient. Smaller (linear) costs are associated with updating u and solving the proximal operator.