A compendium of IRLS algorithms

Bradley Worley

July 20, 2019

Abstract

These notes sketch the derivations of all solvers implemented within the `irls-sandbox` repository.

1 Notation and definitions

$\mathcal{N}\big(\mu, \sigma^2\big)$	Univariate normal distribution with mean μ and variance σ^2
$\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	d-variate normal distribution with means $\boldsymbol{\mu}$ and covariances $\boldsymbol{\Sigma}$
$\mathcal{IG}(\alpha, \beta)$	Inverse-gamma distribution with shape α and scale β
$\mathcal{IN}(\nu, \lambda)$	Inverse Gaussian distribution with mean ν and shape λ

A random variable $x \sim \mathcal{N}\big(\mu, \sigma^2\big)$ has density:

$$f_{\mathcal{N}}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

A random variable $\boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has density:

$$f_{\mathcal{N}_d}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

A random variable $x \sim \mathcal{IG}(\alpha, \beta)$ has density:

$$f_{\mathcal{IG}}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \big(x^{-1}\big)^{\alpha+1} \exp\left\{ -\beta x^{-1} \right\}$$

A random variable $x \sim \mathcal{IN}(\nu, \lambda)$ has density:

$$f_{\mathcal{IN}}(x; \nu, \lambda) = (2\pi)^{-\frac{1}{2}} \lambda^{\frac{3}{2}} x^{-\frac{3}{2}} \exp\left\{ -\frac{\lambda(x-\nu)^2}{2\nu^2 x} \right\}$$

2 Base probabilistic construction

We begin with the measurement model,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{y}, \boldsymbol{\epsilon} \in \mathcal{R}^m$, $\boldsymbol{x} \in \mathcal{R}^n$, and $\boldsymbol{A} \in \mathcal{R}^{m \times n}$.

Assuming $\forall j : \epsilon_j \mid \tau \sim \mathcal{N}(0, \tau^{-1})$ yields the likelihood:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \tau) = (2\pi)^{-\frac{m}{2}} \tau^{\frac{m}{2}} \exp\left\{-\frac{\tau}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2\right\} \tag{2.2}$$

Assuming $\forall i : x_i \mid w_i \sim \mathcal{N}(0, w_i^{-1})$ yields:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{w}) &= \prod_{i=1}^{n} p(x_i \mid w_i) \\ p(x_i \mid w_i) &= (2\pi)^{-\frac{1}{2}} w_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2} w_i |x_i|^2\right\} \end{aligned} \tag{2.3}$$

Assuming $\forall i : w_i \mid \xi \sim \mathcal{IG}(1, \xi/2)$ yields:

$$\begin{aligned} p(\boldsymbol{w} \mid \xi) &= \prod_{i=1}^{n} p(w_i \mid \xi) \\ p(w_i \mid \xi) &= \frac{\xi}{2} w_i^{-2} \exp\left\{-\frac{\xi}{2} w_i^{-1}\right\} \end{aligned} \tag{2.4}$$

Assuming $\xi \sim \mathcal{IG}(n + 1/2, \beta_\xi/2)$ yields:

$$p(\xi) = \left(\frac{\beta_\xi}{2}\right)^{n+\frac{1}{2}} \Gamma\left(n + \frac{1}{2}\right)^{-1} \xi^{-n-\frac{3}{2}} \exp\left\{-\frac{\beta_\xi}{2} \xi^{-1}\right\} \tag{2.5}$$

Assuming $\tau \sim \mathcal{IG}((m + 1)/2, \beta_\tau/2)$ yields:

$$p(\tau) = \left(\frac{\beta_\tau}{2}\right)^{\frac{m+1}{2}} \Gamma\left(\frac{m+1}{2}\right)^{-1} \tau^{-\frac{m+3}{2}} \exp\left\{-\frac{\beta_\tau}{2} \tau^{-1}\right\} \tag{2.6}$$

The final joint distribution is given by,

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \xi, \tau) = \underbrace{p(\boldsymbol{y} \mid \boldsymbol{x}, \tau)\, p(\boldsymbol{x} \mid \boldsymbol{w})\, p(\boldsymbol{w} \mid \xi)}_{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w} \mid \xi, \tau)}\, p(\xi)\, p(\tau) \tag{2.7}$$

3 Equality-constrained IRLS

Define the optimization problem as follows:

$$\begin{aligned}
&\underset{\boldsymbol{x},\boldsymbol{w}}{\text{minimize}}\ f(\boldsymbol{x},\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\big(w_i|x_i|^2 + w_i^{-1}\big) \\
&\text{subject to } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}
\end{aligned} \tag{3.1}$$

3.1 Updates to $\boldsymbol{x}$

3.1.1 Direct

To update $\boldsymbol{x}$, we define $\boldsymbol{W} : W_{ij} = \delta(i-j)w_i$ and form the Lagrangian,

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\lambda}^\top(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \tag{3.2}$$

Setting the $\boldsymbol{x}$-gradient equal to zero,

$$\begin{aligned}
\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) &= \boldsymbol{W}\boldsymbol{x} - \boldsymbol{A}^\top\boldsymbol{\lambda} \\
\implies \boldsymbol{x} &= \boldsymbol{W}^{-1}\boldsymbol{A}^\top\boldsymbol{\lambda}
\end{aligned} \tag{3.3}$$

which leads to the dual objective,

$$g(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^\top\boldsymbol{y} - \frac{1}{2}\boldsymbol{\lambda}^\top\boldsymbol{A}\boldsymbol{W}^{-1}\boldsymbol{A}^\top\boldsymbol{\lambda} \tag{3.4}$$

Maximizing the dual leads to the final value of the updated $\boldsymbol{x}$:

$$\begin{aligned}
&\nabla_{\boldsymbol{\lambda}}g(\boldsymbol{\lambda}) = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{W}^{-1}\boldsymbol{A}^\top\boldsymbol{\lambda} \\
&\implies \boldsymbol{\lambda} = \left(\boldsymbol{A}\boldsymbol{W}^{-1}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{y} \\
&\implies \boldsymbol{x} = \boldsymbol{W}^{-1}\boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{W}^{-1}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{y}
\end{aligned} \tag{3.5}$$

3.1.2 Dual ascent

To avoid matrix inversion, we can perform dual gradient ascent to find $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \kappa\nabla_{\boldsymbol{\lambda}}g(\boldsymbol{\lambda}^t) \tag{3.6}$$

where the step size $\kappa = \min(\boldsymbol{w})$ is an approximation of the optimal step size $\kappa^* = L_g^{-1}$, where

$$L_g = 2\max\operatorname{eig}\big(\nabla^2_{\boldsymbol{\lambda}}g\big) = \max\operatorname{eig}\Big(\boldsymbol{A}\boldsymbol{W}^{-1}\boldsymbol{A}^\top\Big) \approx \max_i w_i^{-1} = 1/\min_i w_i \tag{3.7}$$

3.2 Updates to $\boldsymbol{w}$

Updating each w_i by setting partial derivatives equal to zero yields,

$$\begin{aligned} \forall i : \partial_{w_i} f(\boldsymbol{x}, \boldsymbol{w}) &= \frac{1}{2}|x_i|^2 - \frac{1}{2}w_i^{-2} \\ \implies w_i &= |x_i|^{-1} \end{aligned} \tag{3.8}$$

which is the uncorrected IRLS weight update. In practice, the corrected weight update $w_i = (|x_i|^2 + \zeta)^{-\frac{1}{2}}$, where $\zeta > 0$, is used.

4 Inequality-constrained IRLS

Define the optimization problem as follows:

$$
\begin{aligned}
&\underset{\boldsymbol{x},\boldsymbol{w}}{\text{minimize}}\ f(\boldsymbol{x},\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\big(w_i|x_i|^2 + w_i^{-1}\big) \\
&\text{subject to } \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2 \le c
\end{aligned}
\tag{4.1}
$$

4.1 Updates to $\boldsymbol{x}$

To update $\boldsymbol{x}$, we define $\boldsymbol{W} : W_{ij} = \delta(i-j)w_i$ and $h(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}$, and form the Lagrangian,

$$
\mathcal{L}(\boldsymbol{x},\lambda) = \frac{1}{2}h(\boldsymbol{x}) + \lambda\big(\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 - c^2\big)
\tag{4.2}
$$

Applying the following bound for functions h with L-Lipschitz continuous gradients:

$$
h(\boldsymbol{x}) \le h(\boldsymbol{z}) + (\boldsymbol{x}-\boldsymbol{z})^\top \nabla_{\boldsymbol{x}} h(\boldsymbol{z}) + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2
\tag{4.3}
$$

yields a majorizing function of the Lagrangian,

$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{x},\lambda) = &\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{W}\boldsymbol{z} + (\boldsymbol{x}-\boldsymbol{z})^\top \boldsymbol{W}\boldsymbol{z} + \frac{L}{4}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 \\
&+ \lambda\big(\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 - c^2\big)
\end{aligned}
\tag{4.4}
$$

Setting the $\boldsymbol{x}$-gradient equal to zero yields the primal solution,

$$
\begin{aligned}
\nabla_{\boldsymbol{x}}\mathcal{L}_{\boldsymbol{z}}(\boldsymbol{x},\lambda) &= \boldsymbol{W}\boldsymbol{z} + \frac{L}{2}(\boldsymbol{x}-\boldsymbol{z}) + 2\lambda\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} - 2\lambda\boldsymbol{A}^\top\boldsymbol{y} \\
\implies \boldsymbol{x} &= \frac{2}{L}\left(\boldsymbol{I} - \frac{4\lambda}{4\lambda+L}\boldsymbol{A}^\top\boldsymbol{A}\right)\left(\frac{L}{2}\boldsymbol{z} - \boldsymbol{W}\boldsymbol{z} + 2\lambda\boldsymbol{A}^\top\boldsymbol{y}\right)
\end{aligned}
\tag{4.5}
$$

which leads to residuals of the form,

$$
\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} = \alpha\boldsymbol{r}
\tag{4.6}
$$

where,

$$
\begin{aligned}
\alpha &= \frac{L}{L+4\lambda} \\
\boldsymbol{r} &= \boldsymbol{y} - \boldsymbol{A}\left(\boldsymbol{I} - \frac{2}{L}\boldsymbol{W}\right)\boldsymbol{z}
\end{aligned}
\tag{4.7}
$$

Substituting this result into the primal feasibility condition yields,

$$
\begin{aligned}
&\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 - c^2 \le 0 \\
\implies &\lambda = \max\left\{0, \frac{L}{4}(\|\boldsymbol{r}\|_2/c - 1)\right\}
\end{aligned}
\tag{4.8}
$$

4.2 Updates to $\boldsymbol{w}$

The updates to each w_i are identical to those in equality-constrained IRLS (cf. 3). It appears that this method requires more iterations than e.g. IRLC-EC to converge. Modifying the weights to the “MAP-form” seems to increase convergence rate.

5 Unconstrained IRLS *via* MAP

Define the optimization problem as follows:

$$\underset{\boldsymbol{x},\boldsymbol{w}}{\text{minimize}}\; f(\boldsymbol{x},\boldsymbol{w}) \tag{5.1}$$

where the objective is equal to the negative log-posterior up to a constant scale and shift:

$$\begin{aligned} f(\boldsymbol{x},\boldsymbol{w}) &= -\ln p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{w} \mid \xi,\tau) \\ &= \frac{\tau}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{1}{2}\sum_{i=1}^{n}\big(w_i|x_i|^2 + \xi w_i^{-1} + 3\ln w_i\big) \end{aligned} \tag{5.2}$$

5.1 Updates to $\boldsymbol{x}$

5.1.1 Direct

Fixing $\boldsymbol{w}$, defining $\boldsymbol{W} : W_{ij} = \delta(i-j)w_i$, and setting the $\boldsymbol{x}$-gradient of f equal to zero yields,

$$\begin{aligned} \nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{w}) &= \tau\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} - \tau\boldsymbol{A}^\top\boldsymbol{y} + \boldsymbol{W}\boldsymbol{x} \\ \implies \boldsymbol{x} &= \tau\Big(\boldsymbol{W}+\tau\boldsymbol{A}^\top\boldsymbol{A}\Big)^{-1}\boldsymbol{A}^\top\boldsymbol{y} \end{aligned} \tag{5.3}$$

which is identical to the $\boldsymbol{x}$ update in 6.

5.1.2 Majorize-minimization

Bounding the ℓ_2-norm term leads to an objective that majorizes f,

$$\begin{aligned} f_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{w}) = \frac{\tau}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{z}\|_2^2 + \tau(\boldsymbol{x}-\boldsymbol{z})^\top\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}) + \frac{L\tau}{4}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 \\ + \frac{1}{2}\sum_{i=1}^{n}\big(w_i|x_i|^2 + \xi w_i^{-1} + 3\ln w_i\big) \end{aligned} \tag{5.4}$$

Setting the $\boldsymbol{x}$-gradient of this majorizing function equal to zero yields,

$$\begin{aligned} \nabla_{\boldsymbol{x}} f_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{w}) &= \tau\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}) + \frac{L\tau}{2}(\boldsymbol{x}-\boldsymbol{z}) + \boldsymbol{W}\boldsymbol{x} \\ \implies \boldsymbol{x} &= \left(\frac{L\tau}{2}\boldsymbol{I}+\boldsymbol{W}\right)^{-1}\left(\frac{L\tau}{2}\boldsymbol{z} - \tau\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y})\right) \end{aligned} \tag{5.5}$$

5.2 Updates to $\boldsymbol{w}$

Fixing $\boldsymbol{x}$ and updating each w_i by setting partial derivatives equal to zero yields,

$$\begin{aligned} \forall i : \partial_{w_i} f(\boldsymbol{x},\boldsymbol{w}) &= \frac{1}{2}|x_i|^2 - \frac{\xi}{2}w_i^{-2} + \frac{3}{2}w_i^{-1} \\ \implies w_i &= \frac{\sqrt{4\xi|x_i|^2+9}-3}{2|x_i|^2} \end{aligned} \tag{5.6}$$

which is unique, and implies that the canonical uncorrected IRLS weight update does not maximize the joint posterior distribution $p(\boldsymbol{x}, \boldsymbol{w} \mid \boldsymbol{y}, \xi, \tau)$.

While this update is undefined at $x_i = 0$, its limit is well-defined:

$$\lim_{x_i \to 0} w_i = \xi/3 \tag{5.7}$$

making it prudent to employ a *non-singular* update of the form:

$$w_i = \begin{cases} \frac{\sqrt{4\xi|x_i|^2+9}-3}{2|x_i|^2} & \text{if } |x_i| \neq 0 \\ \frac{\xi}{3} & \text{if } |x_i| = 0 \end{cases} \tag{5.8}$$

6 Unconstrained IRLS *via* EM

Define the optimization problem as follows:

$$\underset{\boldsymbol{x}, q \in \mathcal{Q}}{\text{minimize}}\; F(\boldsymbol{x}, q) \tag{6.1}$$

where F is a partial variational free energy,

$$F(\boldsymbol{x}, q) = -\mathbb{E}_{q(\boldsymbol{w})}[\ln p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w} \mid \xi, \tau)] - \mathbb{H}[q(\boldsymbol{w})] \tag{6.2}$$

and $\mathcal{Q}$ is the set of all probability distributions over $\boldsymbol{w}$.

6.1 Expectation step

In the expectation step, we minimize F with respect to q. The minimum of F is obtained when q is equal to the product of complete conditionals on w_i (cf. 9), i.e.:

$$q(\boldsymbol{w}) = \prod_{i=1}^{n} p(w_i \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}_{-i}, \xi, \tau) \tag{6.3}$$

which can be made apparent by substituting the log-joint into F,

$$F(\boldsymbol{x}, q) = \frac{\tau}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_q\big[w_i|x_i|^2 + \xi w_i^{-1} + 3\ln w_i\big] - \mathbb{H}[q(\boldsymbol{w})] \tag{6.4}$$

and noting that this function is minimized with respect to q when its entropy has the same form as the inner sum of expectations.

Therefore, we may then write the free energy as a function:

$$\begin{aligned} F(\boldsymbol{x}, \boldsymbol{\nu}_{\boldsymbol{w}}, \boldsymbol{\lambda}_{\boldsymbol{w}}) &= \frac{\tau}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{1}{2}\sum_{i=1}^{n} \nu_{w,i}|x_i|^2 \\ &\quad + \frac{\xi}{2}\sum_{i=1}^{n}\big(\nu_{w,i}^{-1} + \lambda_{w,i}^{-1}\big) + \frac{1}{2}\sum_{i=1}^{n}\ln\lambda_{w,i} \end{aligned} \tag{6.5}$$

which can be trivially minimized with respect to $\boldsymbol{\lambda}_{\boldsymbol{w}}$,

$$\begin{aligned} F(\boldsymbol{x}, \boldsymbol{\nu}_{\boldsymbol{w}}) &= \inf_{\boldsymbol{\lambda}_{\boldsymbol{w}}} F(\boldsymbol{x}, \boldsymbol{\nu}_{\boldsymbol{w}}, \boldsymbol{\lambda}_{\boldsymbol{w}}) \\ &= \frac{\tau}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{1}{2}\sum_{i=1}^{n}\big(\nu_{w,i}|x_i|^2 + \xi\nu_{w,i}^{-1}\big) \end{aligned} \tag{6.6}$$

Setting the $\boldsymbol{\nu}_{\boldsymbol{w}}$-gradient of F equal to zero yields,

$$\begin{aligned} \forall i : \partial_{\nu_{w,i}} F(\boldsymbol{x}, \boldsymbol{\nu}_{\boldsymbol{w}}) &= \frac{1}{2}|x_i|^2 - \frac{\xi}{2}\nu_{w,i}^{-2} \\ \implies \nu_{w,i} &= \sqrt{\xi(|x_i|^2)^{-1}} \end{aligned} \tag{6.7}$$

which is equivalent to the uncorrected IRLS update when $\xi = 1$.

6.2 Maximization step

Defining $\boldsymbol{V} : V_{ij} = \delta(i-j)\nu_{w,i}$ and setting the $\boldsymbol{x}$-gradient of F equal to zero yields,

$$\begin{aligned}\nabla_{\boldsymbol{x}} F(\boldsymbol{x}, \boldsymbol{\nu_w}) &= \tau \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \tau \boldsymbol{A}^\top \boldsymbol{y} + \boldsymbol{V}\boldsymbol{x} \\ \implies \boldsymbol{x} &= \tau \Big(\boldsymbol{V} + \tau \boldsymbol{A}^\top \boldsymbol{A}\Big)^{-1} \boldsymbol{A}^\top \boldsymbol{y}\end{aligned} \tag{6.8}$$

which is equivalent to the unconstrained IRLS update for MAP (cf. 5), and admits the same majorize-minimization update scheme.

7 Unconstrained VRLS

Define the optimization problem as follows:

$$\underset{\boldsymbol{\eta}}{\text{minimize}}\ F(\boldsymbol{\eta}) \tag{7.1}$$

where F is the $\boldsymbol{\lambda_w}$-minimized variational free energy with fixed (ξ, τ),

$$\begin{aligned} F(\boldsymbol{\eta}) &= \inf_{\boldsymbol{\lambda_w}} \left\{ -\mathbb{E}_{q(\boldsymbol{x},\boldsymbol{w})}[\ln p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w} \mid \xi, \tau)] - \mathbb{H}[q(\boldsymbol{x}, \boldsymbol{w})] \right\} \\ &= \frac{\tau}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}\|_2^2 + \frac{\tau}{2}\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{\Gamma}) - \frac{1}{2}\ln\det(\boldsymbol{\Gamma}) \\ &\quad + \frac{1}{2}\sum_{i=1}^{n} \nu_{w,i}\left(|\mu_i|^2 + \Gamma_{ii}\right) + \frac{\xi}{2}\sum_{i=1}^{n}\nu_{w,i}^{-1} \end{aligned} \tag{7.2}$$

and $\boldsymbol{\eta} = \{\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\nu_w}\}$ denotes the set of nontrivial variational parameters.

7.1 Updates to $\boldsymbol{\mu}$

7.1.1 Direct

Defining $\boldsymbol{V} : V_{ij} = \delta(i-j)\nu_{w,i}$ and setting the $\boldsymbol{\mu}$-gradient of F to zero yields,

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\eta}) &= \tau \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{\mu} - \tau \boldsymbol{A}^\top \boldsymbol{y} + \boldsymbol{V}\boldsymbol{\mu} \\ \implies \boldsymbol{\mu} &= \tau\left(\boldsymbol{V} + \tau\boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{y} \end{aligned} \tag{7.3}$$

This objective also admits a majorize-minimization update scheme.

7.1.2 Majorize-minimization

Bounding the ℓ_2-norm around $\boldsymbol{z}$ yields,

$$F_{\boldsymbol{z}}(\boldsymbol{\eta}) = \tau\boldsymbol{\mu}^\top\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y}) + \frac{L\tau}{4}\|\boldsymbol{\mu} - \boldsymbol{z}\|_2^2 + \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{V}\boldsymbol{\mu} + \text{const.} \tag{7.4}$$

Setting the $\boldsymbol{\mu}$-gradient equal to zero yields,

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} F_{\boldsymbol{z}}(\boldsymbol{\eta}) &= \tau\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y}) + \frac{L\tau}{2}(\boldsymbol{\mu} - \boldsymbol{z}) + \boldsymbol{V}\boldsymbol{\mu} \\ \implies \boldsymbol{\mu} &= \left(\frac{L\tau}{2}\boldsymbol{I} + \boldsymbol{V}\right)^{-1}\left(\frac{L\tau}{2}\boldsymbol{z} - \tau\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y})\right) \end{aligned} \tag{7.5}$$

7.2 Updates to $\boldsymbol{\Gamma}$

7.2.1 Dense updates

Setting the $\boldsymbol{\Gamma}$-gradient of F to zero yields,

$$\begin{aligned} \nabla_{\boldsymbol{\Gamma}} F(\boldsymbol{\eta}) &= \frac{\tau}{2}\boldsymbol{A}^\top\boldsymbol{A} - \frac{1}{2}\boldsymbol{\Gamma}^{-1} + \frac{1}{2}\boldsymbol{V} \\ \implies \boldsymbol{\Gamma} &= \left(\boldsymbol{V} + \tau\boldsymbol{A}^\top\boldsymbol{A}\right)^{-1} \end{aligned} \tag{7.6}$$

7.2.2 Diagonal updates

Restricting $\boldsymbol{\Gamma} : \Gamma_{ij\cdot} = \delta(i-j)\gamma_i$ and minimizing yields,

$$\begin{aligned} \forall i : \partial_{\gamma_i} F(\boldsymbol{\eta}) &= \frac{\tau}{2}\delta_i - \frac{1}{2}\gamma_i^{-1} + \frac{1}{2}\nu_{w,i} \\ \implies \gamma_i &= (\nu_{w,i} + \tau\delta_i)^{-1} \end{aligned} \tag{7.7}$$

where $\boldsymbol{\delta} \triangleq \mathrm{diag}(\boldsymbol{A}^\top \boldsymbol{A})$.

7.3 Updates to $\boldsymbol{\nu_w}$

Setting the $\boldsymbol{\nu_w}$-gradient of F to zero yields,

$$\begin{aligned} \forall i : \partial_{\nu_{w,i}} F(\boldsymbol{\eta}) &= \frac{1}{2}\big(|\mu_i|^2 + \Gamma_{ii}\big) - \frac{\xi}{2}\nu_{w,i}^{-2} \\ \implies \nu_{w,i} &= \sqrt{\xi(|\mu_i|^2 + \Gamma_{ii})^{-1}} \end{aligned} \tag{7.8}$$

8 Unconstrained VRLS (extended)

Define the optimization problem as follows:

$$\underset{\boldsymbol{\eta}}{\text{minimize}}\ F(\boldsymbol{\eta}) \tag{8.1}$$

where F is the $(\boldsymbol{\lambda}_{\boldsymbol{w}}, \lambda_\xi, \lambda_\tau)$-minimized variational free energy,

$$\begin{aligned}
F(\boldsymbol{\eta}) &= \inf_{\boldsymbol{\lambda}_{\boldsymbol{w}}, \lambda_\xi, \lambda_\tau} \left\{ -\mathbb{E}_{q(\boldsymbol{x}, \boldsymbol{w}, \xi, \tau)}[\ln p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \xi, \tau)] - \mathbb{H}[q(\boldsymbol{x}, \boldsymbol{w}, \xi, \tau)] \right\} \\
&= \frac{\nu_\tau}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}\|_2^2 + \frac{\nu_\tau}{2} \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{\Gamma}) - \frac{1}{2} \ln \det(\boldsymbol{\Gamma}) \\
&\quad + \frac{1}{2} \sum_{i=1}^{n} \nu_{w,i} \left(|\mu_i|^2 + \Gamma_{ii} \right) + \frac{\nu_\xi}{2} \sum_{i=1}^{n} \nu_{w,i}^{-1} \\
&\quad + \frac{\beta_\xi}{2} \nu_\xi^{-1} + \frac{\beta_\tau}{2} \nu_\tau^{-1}
\end{aligned} \tag{8.2}$$

and $\boldsymbol{\eta} = \{\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\nu}_{\boldsymbol{w}}, \nu_\xi, \nu_\tau\}$ denotes the set of nontrivial variational parameters.

8.1 Updates to $\boldsymbol{\mu}$

8.1.1 Direct

Fix me!

8.1.2 Majorize-minimization

Fix me!

8.2 Updates to $\boldsymbol{\Gamma}$

8.2.1 Dense updates

Fix me!

8.2.2 Diagonal updates

Fix me!

8.3 Updates to $\boldsymbol{\nu}_{\boldsymbol{w}}$

Fix me!

8.4 Updates to ν_τ

Fix me!

8.5 Updates to ν_ξ

Fix me!

9 Gibbs sampling

9.1 Complete conditional for $\boldsymbol{x}$

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{w}, \xi, \tau) \propto \exp\left\{-\frac{\tau}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 - \frac{1}{2}\sum_{i=1}^{n} w_i |x_i|^2\right\} \tag{9.1}$$

which is $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ with parameters:

$$\begin{aligned} \boldsymbol{\mu} &= \tau \boldsymbol{\Gamma} \boldsymbol{A}^\top \boldsymbol{y} \\ \boldsymbol{\Gamma} &= \left(\boldsymbol{W} + \tau \boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \end{aligned} \tag{9.2}$$

9.2 Complete conditional for $\boldsymbol{w}$

$$p(w_i \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}_{-i}, \xi, \tau) \propto w_i^{-\frac{3}{2}} \exp\left\{-\frac{1}{2} w_i |x_i|^2 - \frac{\xi}{2} w_i^{-1}\right\} \tag{9.3}$$

which is $\mathcal{IN}(\nu_{w,i}, \lambda_{w,i})$ with parameters:

$$\begin{aligned} \nu_{w,i} &= \sqrt{\frac{\xi}{|x_i|^2}} \\ \lambda_{w,i} &= \xi \end{aligned} \tag{9.4}$$

9.3 Complete conditional for ξ

$$p(\xi \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \tau) \propto \xi^{-\frac{3}{2}} \exp\left\{-\frac{\xi}{2}\sum_{i=1}^{n} w_i^{-1} - \frac{\beta_\xi}{2}\xi^{-1}\right\} \tag{9.5}$$

which is $\mathcal{IN}(\nu_\xi, \lambda_\xi)$ with parameters:

$$\begin{aligned} \nu_\xi &= \sqrt{\beta_\xi \left(\sum_{i=1}^{n} w_i^{-1}\right)^{-1}} \\ \lambda_\xi &= \beta_\xi \end{aligned} \tag{9.6}$$

9.4 Complete conditional for τ

$$p(\tau \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \xi) \propto \tau^{-\frac{3}{2}} \exp\left\{-\frac{\tau}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 - \frac{\beta_\tau}{2}\tau^{-1}\right\} \tag{9.7}$$

which is $\mathcal{IN}(\nu_\tau, \lambda_\tau)$ with parameters,

$$\begin{aligned} \nu_\tau &= \sqrt{\beta_\tau (\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2)^{-1}} \\ \lambda_\tau &= \beta_\tau \end{aligned} \tag{9.8}$$

A Things that eventually need to be rewritten

$$
\begin{aligned}
F &= \mathbb{E}_{q(\boldsymbol{x},\boldsymbol{w},\xi,\tau)}[-\ln p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{w},\xi,\tau)] - \mathbb{H}[q(\boldsymbol{x},\boldsymbol{w},\xi,\tau)] \\
&= \frac{\nu_\tau}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\|_2^2 + \frac{\nu_\tau}{2}\operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{\Gamma}) \\
&\quad + \frac{1}{2}\sum_{i=1}^n \nu_{w,i}\big(|\mu_i|^2+\Gamma_{ii}\big) \\
&\quad + \frac{\nu_\xi}{2}\sum_{i=1}^n\big(\nu_{w,i}^{-1}+\lambda_{w,i}^{-1}\big) + \frac{3}{2}\sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{w})}[\ln w_i] \\
&\quad + \frac{\beta_\xi}{2}\left(\nu_\xi^{-1}+\lambda_\xi^{-1}\right) + \frac{3}{2}\mathbb{E}_{q(\xi)}[\ln \xi] \\
&\quad + \frac{\beta_\tau}{2}\big(\nu_\tau^{-1}+\lambda_\tau^{-1}\big) + \frac{3}{2}\mathbb{E}_{q(\tau)}[\ln \tau] \\
&\quad + \frac{1}{2}\sum_{i=1}^n \ln\lambda_{w,i} + \frac{1}{2}\ln\lambda_\xi + \frac{1}{2}\ln\lambda_\tau \\
&\quad - \frac{1}{2}\ln\det(\boldsymbol{\Gamma})
\end{aligned}
$$

The expected log-precision terms (e.g. $\mathbb{E}_{q(\xi)}[\ln \xi]$) can be written in terms of derivatives of modified Bessel functions, but have no analytic forms. Rather graciously, they cancel from F due to,

$$
\mathbb{H}[f_{\mathcal{IN}}(x;\nu,\lambda)] = \frac{1}{2} - \frac{1}{2}\log\lambda + \frac{3}{2}\mathbb{E}[\ln x] \tag{A.1}
$$

So we simply drop these terms and add some finishing touches:

$$
\begin{aligned}
F(\boldsymbol{\mu},\boldsymbol{\Gamma},\boldsymbol{\nu_w},\nu_\xi,\nu_\tau) = \nu_\tau\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}\|_2^2 &+ \nu_\tau\operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{\Gamma}) \\
&+ \sum_{i=1}^n \nu_{w,i}\big(|\mu_i|^2+\Gamma_{ii}\big) + \nu_\xi\sum_{i=1}^n \nu_{w,i}^{-1} \\
&+ \beta_\xi\nu_\xi^{-1} + \beta_\tau\nu_\tau^{-1} - \ln\det(\boldsymbol{\Gamma})
\end{aligned}
$$

Let $V_{ij} = \delta(i-j)\nu_{w,i}$. Updating $\boldsymbol{\mu}$ yields:

$$
\boldsymbol{\mu} = \nu_\tau\Big(\boldsymbol{V}+\nu_\tau\boldsymbol{A}^\top\boldsymbol{A}\Big)^{-1}\boldsymbol{A}^\top\boldsymbol{y} \tag{A.2}
$$

Updating $\boldsymbol{\Gamma}$ (resp. $\boldsymbol{\gamma}$) yields:

$$
\begin{aligned}
\boldsymbol{\Gamma} &= \Big(\boldsymbol{V}+\nu_\tau\boldsymbol{A}^\top\boldsymbol{A}\Big)^{-1} \\
\forall i : \gamma_i &= (\nu_{w,i}+\nu_\tau\delta_i)^{-1}
\end{aligned} \tag{A.3}
$$

Updating $\nu_{w,i}$ yields:

$$\nu_{w,i} = \sqrt{\nu_\xi (|\mu_i|^2 + \Gamma_{ii})^{-1}} \tag{A.4}$$

Updating ν_ξ yields:

$$\nu_\xi = \sqrt{\beta_\xi \left(\sum_{i=1}^{n} \nu_{w,i}^{-1} \right)^{-1}} \tag{A.5}$$

Updating ν_τ yields:

$$\nu_\tau = \sqrt{\beta_\tau \Big(\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}\|_2^2 + \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{\Gamma}) \Big)^{-1}} \tag{A.6}$$

As a sanity check, these updates correspond to the complete conditionals produced by Gibbs sampling, with expectations taken in the appropriate way.

One interesting note: fixing,

1. $\nu_\xi = 1$,
2. $\nu_\tau = \lambda$, and
3. $\boldsymbol{\Gamma} = \zeta \boldsymbol{I}$,

reduces VRLS back to IRLS with Daubechies' weight correction, with $\boldsymbol{\mu}$ taking the place of $\boldsymbol{x}$ and $\boldsymbol{\nu_w}$ taking the place of $\boldsymbol{w}$.

In VRLS, the entropy term $\mathbb{H}[q(\boldsymbol{x})]$ behaves like a log-barrier enforcing positivity of the weight corrections, $\forall i : \Gamma_{ii} \geq 0$.